VOLUME 6 March 2025 ISSUE 1

JOURNAL ON EMERGING TECHNOLOGIES

ARTICLES

ARTIFICIAL INTELLIGENCE BIASES

EMILE LOZA DE SILES

Abstract

Artificial intelligence (AI) bias, often called "algorithmic bias," is a central focus of AI law and policy debates worldwide. Those discussions have persisted, and erroneously so, in conceptualizing AI bias as a greatly oversimplified and monolithic phenomenon, however. This treatment suggests that AI bias is capable of a relatively facile governance and regulatory approach. One and done, seems to be the idea.

AI biases, however, are many, complex, and often interoperating in their presentations throughout the AI lifecycle. A more scientific, process-oriented problem-solving approach to AI biases is needed to produce fact-based and actionable understandings with which to craft appropriate and effective AI governance and regulation regimes.

This Article adopts systems and process engineering as its guiding rigor and disaggregates the AI biases problem away from simplistic views of AI bias and toward the actionable discernment of individual AI biases. It profiles the AI biases problem space and its complexities. Drawing upon learnings from cognitive engineering and the ethical technology movement, the Article conceives of AI as a human-machine enterprise with human accountability at its core. It then maps out the lifecycle for that joint enterprise as an organizing framework for AI biases governance and control.

This Article then presents the first comprehensive compendium of fifty AI biases synthesized from the literatures of machine learning, AI, computer science, behavioral economics, statistics, epidemiology, psychology, law, and other disciplines; and it translates and interprets

those informed understandings of AI biases and brings them into the legal literature. The law and the corresponding policy debates, however, have little and, as to the great majority of these AI biases, no experience with or understanding of them. The work of this Article, therefore, is all the more pressing as it offers the first rendering of these AI biases as actionable subjects for the creation and implementation of AU governance, public and private, and for the application and development of AI policy and law and of more factually-grounded legal theory.

Following its systems and process engineering approach, the Article organizes the compendium into a first-ever taxonomy of six AI bias categories based upon the domains within the AI lifecycle that those biases do or may impact and, accordingly, the domains that AI governance efforts should address. The Article identifies the AI biases within each category with definitions, descriptions, and AI use case exemplars. To begin to explain the interoperation of AI biases, the Article follows with a discussion of bias injection and other AI bias mechanisms.

With these contributions toward a thoroughly interdisciplinary and process-contextualized understanding of AI biases, this Article enables better informed and grounded AI policy debates, AI governance efforts, and developments of legal theory and law.

Key Words: artificial intelligence, machine learning, bias, algorithmic, AI, cognitive biases, societal or group biases, data biases, learning biases, model biases, use biases, empirical, human-machine cognitive systems, systems engineering, process and quality control, AI law and policy, AI regulation, AI governance, ethnical technology.

Artificial Intelligence Biases

EMILE LOZA DE SILES[*]

To reach the summit, the journey always starts at the foot of the mountain.[1]

INTRODUCTION

Such a simple truth for us earthbound creatures, and one the logic of which is inescapably obvious: Sometimes those aspiring toward whatever goal they wish to reach need reminding that journeys are stepwise endeavors and that the foundational work, the work on the ground, is a prerequisite to the ascent to the summit.

---

[*] Emile Loza de Siles is Assistant Professor of Law of the University of Hawai'i at Mānoa, William S. Richardson School of Law. She has served since 2019 on the Institute of Electrical and Electronics Engineers' (IEEE's) Artificial Intelligence Policy Committee, part of the IEEE Global Initiative on Ethics of Autonomous and Intelligence Systems. She also has served the IEEE Standards Association's P2863 Organizational Governance of AI working group since 2020. Professor Loza is also Chair-Elect of the Association of American Law Schools' Critical Theories Section, having recently completed three years of service to that learned society's 1300-plus member Section on Minority Groups as Chair-Elect, Chair, and Immediate Past-Chair.

Professor Loza joined the legal academy in 2019 with some twenty years' technology and intellectual property law experience serving Cisco, HP, and numerous other innovators in her firm, Technology Law Group. She also served with the U.S. Department of Commerce, Office of General Counsel and the U.S. Federal Trade Commission (FTC). She clerked for Judge Sérgio A. Gutiérrez, Idaho Court of Appeals, and FTC Commissioner Sheila F. Anthony.

Professor Loza's interdisciplinary and translational scholarship centers on artificial intelligence (AI) and law, emphasizing AI governance and regulation; biases and AI-mediated discrimination; and AI impacts on people, liberty, and the rule of law. She holds a technology undergraduate degree, an MBA from the University of Houston; and a juris doctorate from The George Washington University School of Law, with further cybersecurity and data science graduate certificate and studies from Georgetown University and Harvard University, respectively.

I am honored to have presented on this work at the Third Annual Empirical Research Conference on Standardization at Northwestern University School of Law; the Tenth Annual Conference on Governance of Emerging Technologies and Science at Arizona State University Sandra Day O'Connor College of Law; and the Third Annual Michael A. Olivas Writing Institute at the University of California, Davis School of Law. Great thanks to Frank Pasquale, Woodrow Barfield, Ugo Pagallo, Kevin Johnson, Gary Marchant, Margaret Hu, Emad Yaghmaei, Tania Valdez, Thomas Flesher, Daniel Linna, César Cuauhtémoc García Hernández, Juan Perea, Mark Levin, Richard Chen, Miyoko Petit-Toledo, Carina Prunkl, MJ Petroni, Josh Lee Kok Thong, Carlos Ignacio Gutiérrez, Margaret Kwoka, Derek Gundersen, Cory Lenz, and, *qdep*, Daniel Blackaby. Thanks to Isabelle Constant, Micah Miyasato, Siena Schaar, and Brandon Yahiro for their research assistance; and to the members of this Journal. I dedicate this work to the memory of Allison Dang, *qdep*. To my family, *amor y ánimo*. Contact: eloza@hawaii.edu.

[1] Author (for years).

On the complex topic of artificial intelligence (AI) bias and AI-mediated discrimination and other harms, however, this simple truth is often ignored, if not actively avoided. For too long, there has been a proclivity to focus on the summit without contemplating and doing the intellectual work of the ascent: to opine on AI law and policy without having thoroughly examined and understood the factual foundations required to legitimize and sustain the proffered approaches. Like trying to appear magically at the summit without doing the hard work of climbing the mountain, this flaw affects too much law and policy writing that aims to formulate elegant theories and propose workable solutions to the AI bias problem.[2]

Dr. W. Edwards Deming would have a word: "If you can't describe what you are doing as a process, you don't know what you're doing."[3] The father of the total quality management movement, Dr. Deming and his systems of statistical quality control, organizational knowledge, and management philosophy were and are hugely significant to Japan's global industrial leadership today built from worse than nothing after World War II.[4] Even more than thirty years after his death, the reach and influence of Dr. Deming's systems continues well beyond Japan and industry to organizations of all kinds.[5] Indeed, the Deming Prize is the world's oldest, one of its highest, and its first global award for quality.[6]

Dr. Deming speaks directly to the failure to properly understand biases as quality errors within the complex system of human and machine processes that constitute artificial intelligence. We need to know what we are doing. This Article addresses that need in new and

---

[2] This statement does not cast its critical net overbroadly. Those legal scholars whose work is cited herein have added to the understanding of biases affecting artificial intelligence ("AI").

[3] W. Edwards Deming (undated), *quoted in* Jane K. Winn, *Reports of a Blockchain Revolution in Trade Finance Are Greatly Exaggerated* 1, 18 (draft dated Jan. 27, 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3526521.

[4] *See About Dr. W. Edwards Deming*, THE W. EDWARDS DEMING INSTITUTE (undated), https://deming.org/learn/about-dr-deming/ (last visited Mar. 10, 2025).

[5] *See* Tom Connor, *Understanding How Work Is Done – Deming's Theory of Profound Knowledge*, MEDIUM (Apr. 28, 2019), https://medium.com/10x-curiosity/system-of-profound-knowledge-ce8cd368ca62; *see*, *e.g.*, *Deming Management for Quality (part I): Knowledgology – Theory of Knowledge*, CHARTERED QUALITY INST. (Jan. 12, 2018) (part 1 of 4), https://www.quality.org/knowledge/deming-management-quality-part-1-knowledgology-%E2%80%93-theory-knowledge?_gl=1*1rsvj70*_up*MQ..*_ga*MTcxNTc0NTk0OS4xNzEyMDcwMDIy*_ga_PHZCFXV417*MTcxMjA3MDAyMi4xLjEuMTcxMjA3MDA0NS4wLjAuMA.

[6] *See Deming Prize: How Was the Deming Prize Established*, UNION OF JAPANESE SCIENTISTS & ENGINEERS (undated), https://www.juse.or.jp/deming_en/award/index.html (last visited Mar. 10, 2025); *see also Deming Prize*, WIKIPEDIA (last updated Sept. 19, 2024), https://en.wikipedia.org/wiki/Deming_Prize.

comprehensive ways. It provides its theoretical underpinning and then describes and disaggregates that complex system into its constituent processes across a new and comprehensive lifecycle of AI as a human-machine enterprise. It then starts at the foot of the AI biases mountain on a path to identify and organize fifty AI biases into a taxonomy aligned with that lifecycle. It climbs higher to put those AI biases into interoperating motion by examining some key AI bias mechanisms.

The legal literature, government documents, and press accounts are replete with reports of bias in AI systems, particularly predictive systems of which human beings are the computational subjects. Significant problems irrefutably exist as to AI bias and AI-mediated discrimination and other harms. For societal, structural, and data-related reasons, women, people of color, the poor, the disabled, and other minoritized groups disproportionately bear the brunt of those harms. The second and third order impacts of these AI-mediated harms propagate beyond the AI subject, affecting families, communities, institutions, and on to civil society and the rule of law. That AI biases constitute a significant and structural threat is clear. How to address this threat has been less clear.

Even so, AI systems and uses rocket forward. By 2019, 80% of large global companies were already using AI technologies, leaping up from only 10% just five years earlier.[7] This trend has spread and diversified.[8] For example, 56% of businesses recently surveyed use or plan to use AI in their customer operations and 26% in their recruitment operations.[9] Seventy-three percent use or plan to use AI chatbots.[10] The trend is not limited to the business sector, however. The U.S. federal government recently revealed more than 700 AI uses,[11] and more than one-third of Texas' state agencies are already using AI.[12]

Efforts toward establishing laws and other AI guardrails are underway. The international technology community is diligently working

---

[7] Bhaskar Ghosh et al., *Taking a Systems Approach to Adopting AI*, HARV. BUS. REV. (May 9, 2019), https://hbr.org/2019/05/taking-a-systems-approach-to-adopting-ai.
[8] *See* Katherine Haan, *How Businesses Are Using Artificial Intelligence*, FORBES ADVISOR (Rob Watts, ed., last updated Apr. 24, 2023), https://www.forbes.com/advisor/business/software/ai-in-business/.
[9] *See id.*
[10] *See id.*
[11] Madison Alder & Rebecca Heilweil, *U.S. Government Discloses More than 700 AI Use Cases as Biden Administration Promises Regulation*, FEDSCOOP (Oct. 13, 2023), https://fedscoop.com/u-s-government-discloses-more-than-700-ai-use-cases-as-biden-administration-promises-regulation/.
[12] Keaton Peters, *Texas is Exploring Role of AI in Government*, TEX. TRIB. (Jan. 2, 2024), https://www.texastribune.org/2024/01/02/texas-government-artificial-intelligence/.

toward understanding and developing guidelines for identifying and controlling for AI-mediated harms,[13] including from bias associated with the creation of algorithms.[14] As to legal measures, the European Union ("EU") has newly adopted the Artificial Intelligence Act ("AI Act"), the world's first comprehensive AI regulation.[15] The EU AI Act's bias coverage is primarily limited to "high-risk AI systems"[16] and only as to some data biases in those systems.[17] The AI Act once mentions a cognitive bias, *i.e.*, automation bias,[18] and once indicates a use bias[19] as to those systems, but otherwise does not address AI biases. A few states have begun to adopt AI legislation that addresses some AI bias matters,[20] the

---

[13] *See*, *e.g.*, IEEE STANDARDS ASSOCIATION ("IEEE SA"), *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligence Systems*, ver. II (2017), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

[14] For example, the IEEE SA recently approved and published a technical standard as to bias considerations when creating algorithmic systems. *See* IEEE SA, *IEEE P7003 Standard for Algorithmic Bias Considerations* (Jan. 24, 2025), https://standards.ieee.org/ieee/7003/11357/ (last Jan. 31, 2025). This standard culminates a multi-year effort by the IEEE Computer Society/Software & Systems Engineering Standards Committee WP7003 Algorithmic Bias Working Group, a multistakeholder, multinational, and voluntary group of more than 200 members. *See* PAR Details, P7003, IEEE STDS. ASS'N (Root PAR approved Feb. 17, 2017) (PAR expired Dec. 31, 2024), https://development.standards.ieee.org/myproject-web/public/view.html#pardetail/10827; IEEE P7003 Official Roster (last updated Oct. 24, 2024) (on file with author).

[15] *See* Kim Mackrael & Sam Schechner, *European Lawmakers Pass AI Act, World's First Comprehensive AI Law*, WALL ST. J, (last updated March 13, 2024), https://www.wsj.com/tech/ai/ai-act-passes-european-union-law-regulation-e04ec251; Regulation 2024/1689 of the European Parliament and of the Council of 13 March 2024, Artificial Intelligence Act (texts adopted March 13, 2024) [hereinafter EU AI Act], https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html. The EU AI Act entered into force on August 1, 2024. *See* European Commission Press Release IP/24/4123, 2024 European Artificial Intelligence Act Comes into Force (Aug. 1, 2024), https://ec.europa.eu/commission/presscorner/detail/en/IP_24_4123.

[16] *See* EU AI Act, *supra* note 15, at Art. 6(1)-(2) & Annex III. The adopted text of the EU AI Act mentions "bias" nine Article-relevant times and once more in the context of unbiased implementation efforts by the Member States. *See id.* at Art. 70(1).

[17] *See id.* at Art. 10(2)(f)-(g), (5)-(5)(a), (e), (f) & Annex XI, § 2(c).

[18] *See id.* at Art. 14(4)(b).

[19] *See id.* at Art. 15(4).

[20] *See*, *e.g.*, Cal. S.B. 36, Pretrial Release: Risk Assessment Tools (enacted Oct. 8, 2019), https://leginfo.legislature.ca.gov/faces/billStatusClient.xhtml?bill_id=201920200SB36; Colo. SB21-169, Restrict Insurers' Use Of External Consumer Data (enacted July 6, 2021), https://leg.colorado.gov/bills/sb21-169; Ill. Pub. Act No. 102-0047 (enacted July 9, 2021) (amending Artificial Intelligence Video Interview Act), https://www.ilga.gov/legislation/BillStatus.asp?DocNum=53&GAID=16&DocTypeID=HB&LegId=127865&SessionID=110&GA=102.

United States generally and at the federal level is far behind the EU in this regard.[21]

These technical and legal measures will help. They seem unlikely, however, to provide the scope or granularity needed to develop actionable governance measures with which AI developers and users alike may identify and address the spectrum of relevant biases that affect AI systems and uses in burgeoning, diverse, and fragmented AI markets.[22] Companies and agencies continue to aim, but with more aspiration than substantive action, toward the summit of responsible AI design, development, and usage. A comprehensive wayfinding guide for the ascent through the complex terrain of AI biases is sorely needed.

This Article works to fill that gap and to light the path toward the summit in four parts. It brings a number of new contributions into the AI and law literature, drawing extensively from the technical and other disciplines to translate, interpret, and contextualize, for law, important understandings about AI biases. In Part I, it characterizes and reframes the confounding problem space of AI biases. Among its principal workings, the Article discards the long-prevalent and easier, but erroneous view of "AI bias" as a monolith. Its rationale? That view is

---

Several AI bias-encompassing legislative proposals have failed. *See*, *e.g.*, Mass. H. 4029, An Act Relative to Algorithmic Accountability and Bias Prevention Act (introduced July 29, 2021), https://custom.statenet.com/public/resources.cgi?id=ID:bill:MA2021000H4029&ciq=ncsl&client_md=e5ead5db4678349b6f0ec1a035836555&mode=current_text; *Legislation Related to Artificial Intelligence*, NAT'L CONF. OF STATE LEGISLATURES ("NCSL") (Jan. 31, 2023), https://www.ncsl.org/technology-and-communication/legislation-related-to-artificial-intelligence#anchor10850.

See generally also Artificial Intelligence 2023 Legislation, NCSL (table last updated Jan. 12, 2024), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation; Rachel Wright, *Artificial Intelligence in the States: Emerging Legislation*, COUNCIL OF ST. GOV'S (Dec. 6, 2023), https://www.csg.org/2023/12/06/artificial-intelligence-in-the-states-emerging-legislation/.

[21] This statement pertains to federal legislation. The executive branch was making progress on AI regulation within the administrative state. *See*, *e.g.*, President Joseph R. Biden, Jr., Exec. Ord. No. 14,110, 88 Fed. Reg. 75,191, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (Oct. 30, 2023). That Executive Order was rescinded by the new administration, however, and whether any AI bias-protective measures remain or will be forthcoming is an open question. *See* David Shepardson, *Trump Revokes Biden Executive Order on Addressing AI Risks*, REUTERS (Jan. 21, 2025), https://www.reuters.com/technology/artificial-intelligence/trump-revokes-biden-executive-order-addressing-ai-risks-2025-01-21/.

[22] A likely exception will be IEEE SA's forthcoming P2863 standard for the organizational governance of AI. Once finalized and promulgated, that standard, in recommended practice form, will provide a comprehensive methodology by which to surface, examine, and develop prioritized action plans to address multiple potential AI concerns, including bias. *See* Project Authorization Request ("PAR") 2863, *Recommended Practice for Organizational Governance of Artificial Intelligence*, IEEE STDS. ASS'N (PAR expiring Dec. 31, 2025), https://development.standards.ieee.org/myproject-web/public/view.html#pardetail/7330.

overly simplistic, inadequately fact-informed, and thus inactionable toward the meaningful governance and control of AI biases and the mitigation or elimination of the harms caused thereby.

The Article makes three novel and significant contributions through a law and systems and process engineering approach to that problem space. In Part II, it applies the theory of joint human-machine cognitive systems to AI to reveal and document that humans are integrally involved in all phases of the lifecycle of AI as human-machine enterprise. This rational reality-based, re-conceptualized, and re-peopled AI lifecycle clarifies that humans unavoidably remain at the center of all decisions and actions involving AI systems, including fully autonomous ones. At the levers of AI governance and control, humans must always bear responsibility for actuating those levers, [23] and, although applied to the AI biases problem, this Part has much broader application to and import for AI governance and control more generally. The conceptualization of AI as a human-machine enterprise that operates along the thus-presented AI lifecycle provides the organizing framework for what comes next.

In Part III, the Article disaggregates "AI bias" into a taxonomy of fifty different biases (collectively, "AI biases") organized in six impact domains ("domains"), those biases categorized based upon which aspects of the AI lifecycle they do or may impact. Two of these domains are global, meaning that biases occurring within those domains have the potential to impact across the entire AI lifecycle. The remaining four domains represent specific phases of the AI lifecycle in which the impacts of certain biases there are localized, rather than global. This Article presents and describes each of the taxonomy's six domains; identifies the AI biases grouped within each; and defines and describes at least one exemplar from that group, providing brief background, as necessary, for better understanding.

A systems and process engineering approach provides the organizational rigor for the taxonomy presented in Part III. The Article populates that taxonomy through a dauntless translational and interdisciplinary synthesis across machine learning, statistical, behavioral economics, psychological, legal, and other literature to bring order, discernment, and actionability to the complex field of often intersecting AI biases. That grounding and actionability are what render possible the necessary and long-needed governance and control of AI biases. The AI biases and their impact domains, as theorized within the

[23] *See generally* BRIAN CHRISTIAN, THE ALIGNMENT PROBLEM: MACHINE LEARNING AND HUMAN VALUES (2021).

taxonomy, constitute the moving parts that operate within the context of the organizing framework presented in Part II.

In Part IV, the Article places these AI biases into motion through its descriptions and illustrative animations of AI bias mechanisms: three by which AI biases arise, propagate, and amplify throughout the AI lifecycle; and a fourth focused on intercausality by which AI biases may give rise to each other and otherwise interdigitate, including through the interoperation of two or more of these AI bias mechanisms.

The Article's concluding remarks focus all of the Article's novel work toward two principal objectives: to demystify and rationalize the complex field of facts that are AI biases and the mechanisms by which they arise and interact; and to situate, or rather recognize, the situation of humans as integral to and the prime locus of governance and control in AI, which is and must be a human-machine enterprise. These objectives pursue a single goal: to render informed AI governance and control of AI biases actionable now and so protect people, communities, civil society, and the rule of law from the AI-mediated harms that threaten all.

I. AI BIASES: THE PROBLEM SPACE

Bias is error. [24] This is a foundational statement of fact. [25]

---

[24] From a technical perspective, not all bias is "bad." Inductive bias, for example, is essential to and is intentionally harnessed to achieve an AI model's ability to classify, that is, to discriminate between, two potential computed results to favor the more useful result over another less useful one. *See* Tom M. Mitchell, *The Need for Biases in Learning Generalizations*, RUTGERS UNIV. COMP. SCI. DEP'T, Tech. Rep. No. CBM-TR-117 at 1 (1980); Thomas Hellström, Virginia Dignum & Suna Bencsch, *Bias in Machine Learning: What Is It Good For?* at 1, 2 (last updated Sept. 20, 2020), https://arxiv.org/abs/2004.00686; Emile Loza de Siles, *AI, on the Law of the Elephant: Toward Understanding Artificial Intelligence*, 69 BUFF. L. REV. 1389, 1402-03 (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3682835.

As authority, Hellström and his coauthors point to an encyclopedic reference published by Oxford University Press and that reference's definition of 37 types of bias. *See* Hellström et al., *supra*, at 3 (citing A DICTIONARY OF EPIDEMIOLOGY (Miquel Porta, ed., 5th ed. 2008) [hereinafter 2008 DICTIONARY]). They then winnow out most of those bias types as being irrelevant to machine learning because the 2008 DICTIONARY describes them within medical-scientific contexts as indeed it would, given its epidemiology focus. A reframing and re-examination of these exclusions may be warranted, as well as the use of the latest edition of that well-established reference. *See* A DICTIONARY OF EPIDEMIOLOGY (Miquel Porta, ed., 6th ed. 2014) [hereinafter 2014 DICTIONARY], https://www.oxfordreference.com/display/10.1093/acref/9780199976720.001.0001/acref-9780199976720. The earlier edition contains 37 bias entries, whereas this later contains 149. *Compare* Hellström et al., *supra*, at 3 *with* Search of 2014 DICTIONARY (criterion: "bias") (last visited Aug. 9, 2024).

[25] This usage of bias is not a normative construct or a commentary of what comparative norms should be.

Grounding the AI biases problem space in fact is the first step in the essential foundational work required. Grounding the problem space in this fact is, indeed, the only way to reach the summit of understanding AI biases and how governance and control systems and law and policy may appropriately address those biases.

Bias is error, and errors create results that are inaccurate and imprecise, that is, results that are untrue. The goal of all analysis is to arrive at a true understanding of facts, that is, the truth. Bias, therefore, is a systematic, and not a random, set of one or more errors that produce results that deviate from the truth.[26]

Despite the abundant use of the singular term "AI bias," or often erroneously synonymously, "algorithmic bias," bias is not a singular phenomenon. Multitudes of different biases exist. Oxford University's online CATALOGUE OF BIAS contains many pages listing and explaining many different types of bias.[27] Within the category of cognitive biases alone, a crowdsourced list numbers the types at sixty-one,[28] and another group numbers just the most relevant cognitive biases at 109.[29] Turning more specifically to the AI technology space and its problems aside, the National Institute for Standards and Technology's report on AI biases identifies some forty such biases.[30] Although there are some issues with the report, the agency correctly observes that the problem of AI biases goes beyond data biases.[31]

Bias is a highly complex subject in its own right and even more so in an AI context. There is the question of which biases are relevant to consider, for example. Researchers at Technische Universität München and Siemens AG place the number of cognitive biases at more than two hundred, and they determined that no fewer than thirty-three of them

---

[26] 2014 DICTIONARY, *supra* note 24, at 21.
[27] *See Catalogue of Bias,* CENTRE FOR EVIDENCE-BASED MEDICINE, UNIVERSITY OF OXFORD, (Carl Heneghan & David Nunan eds., 2020) [hereinafter OXFORD BIAS CATALOGUE], https://catalogofbias.org/biases/.
[28] *List of Cognitive Biases*, WIKIPEDIA (last updated Feb. 25, 2024), https://en.wikipedia.org/wiki/List_of_cognitive_biases.
[29] *Cognitive Biases: A List of the Most Relevant Biases in Behavioral Economics*, THE DECISION LAB (undated), https://thedecisionlab.com/biases (last visited Mar. 10, 2025).
[30] *See* REVA SCHWARTZ ET AL., NAT'L INST. OF STANDARDS & TECH., TOWARDS A STANDARD FOR IDENTIFYING AND MANAGING BIAS IN ARTIFICIAL INTELLIGENCE, Spec. Pub. No. 1270 at 49-77 (Mar. 2022) [hereinafter NIST REP'T] (glossary of some 40 biases), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf. *See also* Ninareh Mehrabi et al., *A Survey on Bias and Fairness in Machine Learning* 3–4, ARXIV (2022), https://arxiv.org/pdf/1908.09635.pdf.
[31] *See There's More to AI Bias Than Biased Data, NIST Report Highlights*, NAT'L INST. OF STANDARDS & TECH. (Mar. 16, 2022), https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights.

are essential to consider during the process of architecting software.[32]

Further complicating discernment and comprehension as to biases and much like the term "artificial intelligence" itself, there are no unifying definitions for numerous types of biases.[33] As a result, many biases have multiple monikers. For example, class membership bias, a cognitive bias resulting from drawing certain inferences about a person based upon their membership within a group, or class, is also dubbed something called "ecological fallacy."[34] Moreover, some sources refer synonymously to biases that have similar names, but are not in fact the same thing. "Observational bias," "observer bias," and "observation bias" are cases in point.[35]

In addition, the varying applications of biases drive their differing and context-specific interpretations. For instance, ascertainment bias is a type of selection bias and observer bias that, respectively, introduces errors due to the data or groups that are selected for analysis and, consequently, leads those doing the analyses, *i.e.*, the observers, to arrive at erroneous understandings or conclusions.[36] Ascertainment bias has been discussed in the context of genetic and other medical studies,[37] and those discussions encompass measures for revealing the presence of ascertainment bias, evaluating its impacts, and eliminating or limiting those impacts,[38] what this Article collectively coins "governance control"

---

[32] *See* Akash Manjunath et al., *Decision Making and Cognitive Biases in Designing Software Architectures*, 2018 IEEE Int'l Conf. on Software Architecture Companion (ICSA-C) 1, 11 (2018), doi: 10.1109/ICSA-C.2018.00022.

[33] *See* Alexandra Olteanu et al., *Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries*, 2 FRONT. BIG DATA 1, 2 (Juergen Pfeffer ed., 2019), doi: 10.3389/fdata.2019.00013.

[34] NIST REP'T, *supra* note 30, at 23 & 50.

[35] *See*, *e.g.*, Hellström et al., *supra* note 24, at 3 (observational bias); 2014 DICTIONARY, *supra* note 24, at 149 (observer bias); Jianbing Jin et al., *Machine Learning for Observation Bias Correction with Application to Dust Storm Data Assimilation*, 19 ATMOS. CHEM. PHYS. 10009-10026, at § 1, para. 4 (2019), https://doi.org/10.5194/acp-19-10009-2019 (observation bias). Observer bias, however, is distinguishable from, narrower than, and, indeed, may be a cause of observation bias. By contrast, observational bias may describe bias at the observer level or at the "higher" and potentially more inclusive level of observation.

[36] *See* Elizabeth A. Spencer & Jon A. Brassey, *Ascertainment bias* (2017), *in* OXFORD BIAS CATALOGUE, *supra* note 27, https://catalogofbias.org/biases/ascertainment-bias/.

[37] Ascertainment bias within genetics analyses, for example. refers to the statistically calculated chance associated with the selecting persons from smaller or larger families. *See id.*

[38] *See* D. Michal Freedman & Ruth M. Pfeiffer, *Ascertainment Bias in Statin Use and Alzheimer Disease Incidence*, 74 JAMA NEUROL. 868 (July 1, 2017), doi:

measures.

So, there are existing conversations about ascertainment bias in genetic or medical applications of that concept. How should ascertainment bias, however, be understood, revealed, and controlled for where medical diagnoses, prognoses, or care plans are AI-mediated?[39] What, as a further example, are the implications of ascertainment bias when using genetic data to predict who are the DNA contributors to mixed blood samples found at crime scenes?[40]

If the import of such questions is understood, the answers to them are not. As demonstrated by this small example, translational and interdisciplinary work is required to bring such domain-specific understandings of biases into AI contexts to identify where and how those biases may operate and impact upon the AI systems and uses and, most importantly, upon the computational results to which humans defer or upon which they base their own decisions.

Similarly, and in part due to the lack of a unified language of biases, there is no agreed comprehensive taxonomy within which to frame them. Some sources treat different types of biases as hierarchically related to one another, whereas others do not.[41] The categorization of biases is necessary to simplify the problem space. Without that tool, biases seem like so many mathematical nesting dolls with one bias being a subtype of another or like a Rube Goldberg machine in which the presence of one type of bias triggers others to come into being and join into a disentangleable cascade of biases.

For example, publication bias as to the choices of which court decisions to publish may result in sampling or selection biases that undermine the validity of empirical legal studies.[42] Such publication bias also may result in cognitive biases that impact judicial decision-making, the development and advancement of novel legal theories in litigation,

---

10.1001/jamaneurol.2017.0427. Freedman and Pfeiffer provided comments on an Alzheimer's disease-related study. They pointed out that the validity of the study results may be almost entirely undermined by ascertainment bias caused, in part, by the choice of Medicare claims data that formed the basis of the study's analyses. *See id.* as to Julie M. Zissimopoulos et al., *Sex and Race Differences in the Association Between Statin Use and the Incidence of Alzheimer Disease*, 74 JAMA NEUROL. 225 (Feb. 1, 2017), doi: 10.1001/jamaneurol.2016.3783.

[39] *See*, *e.g.*, Gustavo Rodríguez Leal, *Revolutionizing Healthcare: The Synergy of AI and Genetics*, MEX. BUS. NEWS (Sept. 25, 2023), https://mexicobusiness.news/health/news/revolutionizing-healthcare-synergy-ai-and-genetics.

[40] *See* State v. Pickett, 466 N.J. Super. 270, 246 A.3d 279 (App. Div. 2021).

[41] *Compare* Hellström et al., *supra* note 24, at 3 *with* 2014 DICTIONARY, *supra* note 24, at 149 (information, observation, & measurement biases).

[42] *See* Edward K. Cheng, *Detection and Correction of Case-Publication Bias*, 47 J. LEGAL STUD. 151, 153 (2018); NIST Rep't, *supra* note 30, at 8, 15, 27.

and, ultimately, outcomes for litigants.[43] The relationships between biases are complicated, and efforts to control for one type of bias may be in conflict with those efforts for other biases, requiring tricky judgments in the attempt to balance trade-offs between them.[44]

Biases are errors and so the impacts, actual and potential, of those errors across the AI lifecycle must be illuminated and thoughtfully considered. Whether and, if so, the extent to and modes by which those AI biases are unfair is a foundational inquiry. Here, the problematic perception of AI bias as monolithic again enters into play. Many, if not most, lay, law, and policy discussions, however, treat AI bias categorically as a mechanized type of "unfair bias."[45]

The necessary inquiry, however, should first identify the AI biases in play, including interdigitating with other such biases within particularized factual context. Then, the inquiry should compare the harms or potential harms of those AI biases with some standard against which fairness or unfairness may appropriately adjudged. Just as the law is, perhaps more often than not, a poor substitute for justice, unfairness does not fully correspond to illegality. That said and although its guidance must be AI-contextualized, the law provides some baseline parameters by which to gauge what is generally unfair, such as a violation of the Universal Declaration of Human Rights, to what is specifically unfair, such as discrimination based upon race, gender, age, and disability, for instance, under civil rights laws or as unfair trade practices under consumer protection law. At a minimum, AI biases that result or may result in such harms are unfair.

All of this, then, is the complex problem space of AI biases.

II. AI AS HUMAN-MACHINE ENTERPRISE and ITS LIFECYCLE

As a beginning frame, this Article and other works conceptualize

---

[43] *See* Cheng, *supra* note 42, at 152-56.

[44] *See*, *e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L. J. 2218, 2236-38, 2248-51 (2019); Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-offs in the Fair Determination of Risk Scores*, *in* 8th Innovations in Theoretical Computer Science Conference (2017), *in* 67 LEIBNIZ INT'L PRO. IN INFORMATICS 43:1–43:23 (2018).

[45] *See Ethics Guidelines for Trustworthy AI*. European Commission Independent High-Level Expert Group on Artificial Intelligence, at 36 (Apr. 8, 2019), doi:10.2759/346720. These Guidelines define bias as "an inclination of prejudice towards or against a person, object, or position" and "unfair bias" as "bias that can result in discriminatory and/or unfair outcomes." *Id*.

AI as a human-machine enterprise.[46] The origins of this concept began with David D. Woods and Erik Hollnagel's introduction of the theory of "joint human-machine cognitive systems" in 1983.[47] In this Part, the Article discusses Hollnagel and Woods' theory and then adapts it and explains its rationale for adapting this theory as a basis for AI governance control.

A. Joint Human-Machine Cognitive System Theory

What is a joint human-machine cognitive system? To illustrate, just think of the contacts stored in a smartphone's memory and perhaps backed up to the cloud. The phone's owner likely has chosen not to remember her friends and family's phone numbers, or most of them, anyway. Why is this? Because she has augmented her human memory with the artificial memory resting within her smartphone device and in one or more server farms somewhere in the world. She need not take up mindspace committing all those numbers to her, at times flawed or overburdened, human memory and later recalling them or changes to them as people move or get new numbers. The artificial memory works just fine. She knows how to use the contacts function and to recover those numbers if her phone is damaged or goes missing. The artificial system is set to regularly back up her contacts, and it automatically detects contact entries that are potentially duplicative, in whole or part. It presents her with the opportunity to instruct it to merge contacts to keep the information tidy and organized for easy use. This everyday example illustrates the basic point of just a small part of the system of cognition that jointly interoperates between the mind and memory of the smartphone owner and the user interface and memory of the smartphone.

Some forty-one years ago, Erik Hollnagel of Denmark's Risø National Laboratory and David D. Woods, a Westinghouse researcher, both now professors emeriti, pioneered the field of cognitive systems engineering and wrote their ground-breaking work theorizing "joint

---

[46] *See*, *e.g.*, Loza de Siles, *AI, on the Law of the Elephant*, *supra* note 24, at 1415-16; Gary Marchant, "*Soft Law" Governance of Artificial Intelligence*, AI PULSE 1, 2 (Jan. 25, 2019), https://escholarship.org/uc/item/0jq252ks.

[47] Erik Hollnagel & David D. Woods, *Cognitive Systems Engineering: New Wine in New Bottles*, 18 INT'L J. MAN-MACH. STUD. 583 (1983), https://doi.org/10.1016/S0020-7373(83)80034-0; *see* David D. Woods, *Cognitive Technologies: The Design of Joint Human-Machine Cognitive Systems*, 6 AI MAG. 86 (1985), https://doi.org/10.1609/aimag.v6i4.511.

human-machine cognitive systems."[48] Writing in the Association for the Advancement of Artificial Intelligence' journal, AI MAGAZINE, in 1985, Woods presented and began to socialize the pair's new theory within the AI community.[49] This Article introduces this pioneering theory into the legal literature and explains it as emerged from the fountainhead.

To first level set the terminology, "cognition" is "[t]he action or faculty of knowing taken in its widest sense, including sensation, perception, conception," and so on, and is distinguishable from volition or feeling.[50] A "cognitive system" adaptively uses knowledge about itself and the world and manipulates symbols to model or represent that knowledge; with these, the cognitive system has the capacity to consider the problem or goal before it from various perspectives and thereby plan, modify, and produce "intelligence action," that is, behaviors directed toward the problem or goal.[51] Thus, a cognitive system is both data- and concept-driven.[52]

Human beings are natural cognitive systems.[53] Computational systems that have the capacity to perform tasks normally ascribed to humans also constitute cognitive systems, although human-made artificial ones.[54] Such computational systems fit ideally within the International Organization for Standardization, or ISO, and International Electrotechnical Commission, or IEC, harmonized their definition of "artificial intelligence" as such systems having the capabilities "to perform functions that are generally associated with human intelligence such as reasoning and learning."[55] Applying this current terminology, AI systems constitute artificial cognitive systems.

Understood as separate cognitive systems, the human user of the AI system holds in mind a particular model of what that machine is and how it functions.[56] Likewise, the AI system, having been designed and developed with a particular model of who, generally speaking, its users

---

[48] *See* Hollnagel & Woods, *supra* note 47; *see, e.g.,* Emilie M. Roth et al., *Designing Collaborative Planning Systems: Putting Joint Cognitive Systems Principles to Practice*, *in* COGNITIVE SYSTEMS ENGINEERING: THE FUTURE FOR A CHANGING WORLD 247, 248 (Philip J. Smith & Robert R. Hoffman, eds., 2018).
[49] *See generally* Woods, *supra* note 47.
[50] *Cognition*, OXFORD ENG. DICT. § 2(a) (last updated July 2023), https://www.oed.com/dictionary/cognition_n?tab=meaning_and_use.
[51] *See* Hollnagel & Woods, *supra* note 47, at 589.
[52] *See id.*
[53] *See id.*
[54] *See* Woods, *supra* note 47, at 86.
[55] INT'L ORG. FOR STANDARDIZATION & INT'L ELECTROTECHNICAL COMM., INT'L STANDARD ISO/IEC 2382:2015, *Information Technology – Vocabulary* (2015), https://www.iso.org/standard/63598.html.
[56] *See* Hollnagel & Woods, *supra* note 47, at 589-90.

are and how they will use and interpret the output of the AI system, functions based upon that, one might say, statically stereotypical model of its users.[57]

These individuated model views of the two participants in the cognitive exercise produce the concomitant and discretized view of the human and machine cognitive systems as two independent parts. This discretized view, however, misconceives what is the true reality of and better nature of the interaction between those cognitive systems.

These two systems combine to form a single joint human-machine cognitive system that functions as one with the human intelligence contributing its unique capacities, *e.g.*, judgment, experience, intuition, and the artificial intelligence contributing its own, *e.g.*, pattern recognition within vast quanta of data, rapidity of complex calculations.[58] The whole of these two cognitive systems working together is greater than the sum of two otherwise-viewed independent parts.[59] Where the user's training and interactions with the system and AI system's design and development are directed toward their respective roles within the joint human-machine cognitive system as a whole, a host of problems are foreclosed, including problems that have legal bearing, such as the black-boxing of AI systems and other transparency, accountability, and explainability,[60] which, here, dubbed together as "decision provenance" problems. Under this integrated and role-based cognitive model, the effectiveness of the joint human-machine cognitive system is improved in comparison with other solo[61] or discretized models.[62]

The United States Department of Defense ("Department") follows the Hollnagel-Woods model as it increasingly explores, funds the

---

[57] *See id.*

[58] *See* Woods, *supra* note 47, at 87 & fig. 1; J. E. (Hans) Korteling et al., *Human-versus Artificial Intelligence*, 4 FRONTIERS IN A.I. (Esma Aïmeur ed., 2021), https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.622364/full.

[59] *See* Hollnagel & Woods, *supra* note 47, at 586-87.

[60] *See*, *e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 3 (2015) (discussing "black box" to mean AI or other algorithmic systems that remain mysterious as to their function) ("[W]e can observe its inputs and outputs, but we cannot tell how one becomes the other.").

[61] Artificial cognitive systems operating solo without the cognitive collaboration of humans may present grave and even deadly risks and consequences. *See*, *e.g.*, Joe Hernandez, *A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says*, NAT'L PUB. RADIO (June 1, 2021), https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d.

[62] For example, the human user in the joint human-machine cognitive systems does not cede control of the interaction to the machine and becomes more capable of identifying and rejecting poor quality results that the machine may deliver. *See* Woods, *supra* note 47, at 86; Hollnagel & Woods, *supra* note 47, at 584-85.

development of, and integrates fully- and semi-autonomous weapons, vehicles, and other systems into its defensive and war-fighting capabilities.[63] This model, as the Defense Science Board ("Board") emphasizes, points away from the discretized models view, and attendant autonomy classification systems, as diverting critical "focus from the fact that *all autonomous systems are joint human-machine cognitive systems*[.]"[64] The Board's views echo its agreement with some of the flaws identified by Hollnagel and Woods as inherent in the discretized models view and point to the controlling advantage of the joint system model, stating:

> Treating autonomy as a widget or "black box" supports an "us versus the computer" attitude . . . rather than the more appropriate understanding that *there are no fully autonomous systems just as there are no fully autonomous soldiers, sailors, airmen or Marines*. Perhaps the most important message . . . is that all systems are supervised by humans to some degree, and the best capabilities result from the coordination and collaboration of humans and machines.[65]

Commentators are increasingly calling for the recognition of how humans and AI systems operate collaboratively and thus marking the advancing relevance of the Hollnagel-Woods theory as its adoption, if without attribution, demonstrates.[66]

The power of the Hollnagel-Woods theory of joint human-machine cognitive systems is its accurate representation of the ways in which humans collaboratively engage with and rely upon artificially intelligent "machines." Future works should apply this powerful

---

[63] *See Final Report of the Defense Science Board (DSB) Task Force on the Role of Autonomy in Department of Defense (DoD) Systems* § 3.2, OFF. OF UNDER SECRETARY OF DEF. FOR ACQUISITION, TECH. & LOGISTICS, DEP'T OF DEF., 23 (2012), www.fas.org/irp/agency/dod/dsb/autonomy.pdf [hereinafter DOD]. Woods served on the task force, and his and Hollnagel's work featured in this report. *See id.* at App'x B & D.

Although not citing to Woods and Hollnagel's groundbreaking work, Center for a New American Security executive vice president and former Army Ranger Paul Scharre discusses autonomous weapon systems and the advantages of combining human and machine intelligences into hybridized "cognitive architectures." *See* Paul Scharre, *Centaur Warfighting: The False Choice of Humans vs. Automation*, 30 TEMP. INT'L & COMP. L.J. 151, 152 (2016).

[64] DoD, *supra* note 63, at 24 (emphasis retained).

[65] *Id* (emphasis retained).

[66] *See, e.g.*, H. James Wilson & Paul R. Daugherty, *Collaborative Intelligence: Humans and AI Are Joining Forces,* HARV. BUS. REV. (July-Aug. 2018), https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces.

construct broadly across the legal spectrum and particularly as to human accountability within any AI governance system and the essential requirement that the rule of law must be preserved and strengthened as these joint cognitive systems operate and come on board.

B. Artificial Intelligence as Human-Machine Enterprise

The Article grounds its view of AI governance, whether organizational or regulatory governance, in Hollnagel and Woods' theory of AI as a joint human-machine cognitive system. This view stands for two principles.

First, AI system uses may expedite, render more efficient, augment, and thus improve, human cognition; and, where those assumed improvements are empirically demonstrated, such uses should proceed and those benefits thereby accrue. Second, however, the human role in the AI-as-human-machine-enterprise should never be obviated. Humans are inextricably involved, directly or indirectly, in every aspect of the AI cycle from the conceptualization, design, development, procurement, implementation, and on through use of AI systems (end-to-end collectively, "AI lifecycle"). Therefore, where AI biases threaten or produce discrimination or other harms, humans must always bear responsibility. Consequently, humans have the capacity and the duty to act with intention as to all AI biases.

Does this mean that humans have the capacity to eliminate AI biases altogether? Likely no. It means that humans must not be permitted to ignore or feign ignorance of those biases and their attendant harms and potential harms; or to disclaim their responsibilities to detect and correct or otherwise responsibly address those AI biases. Auditing, component and system validations, and other continuous quality improvement methodologies, by which AI biases may be surfaced for action, are essential. It is by these means that humans become capable of knowing, as they must, which among the many and, in some combinations, interdigitating[67] AI biases impact upon the AI lifecycle; and how and the extent to which those errors produce deviations from the truth in the functioning and use of AI systems and their computed results.

Eyes thus opened, humans then must engage in disciplined decision-making as to whether and, if so, how and with what quantifying and corrective measures a given AI system should be used. This is true

[67] An examination of AI bias mechanisms, including as to how some biases interoperate and amplify others, is reserved for a future work.

and right even where AI systems operate autonomously. At some point, humans are involved in decision-making and action that put those systems into their self-driven flights. In sum, AI biases must remain subjects of human intentionality and assigned as remaining within human control. President Truman's credo is as true today with respect to AI as it ever was with the decisions and crises he faced while in office long ago: "The buck stops here!"[68] The AI buck must always stop with humans.

C. C. AI-as-Human-Machine Enterprise: A New and Actionable Lifecycle Conception

Most depictions of the AI lifecycle are narrow in scope[69] and agranular.[70] They focus upon and prioritize the "machine" parts of AI systems with the "human" parts relegated to minor roles. For example, the lifecycle adopted and approved by the Organisation of Economic Co-operation and Development's ("OECD's") AI Group of Experts depicts humans as being "sensors" or "actuators."[71] As sensors, those humans constitute one mechanism by which data are obtained from the external environment and brought into the computational scope of the AI system.[72] As actuators, those humans constitute one means by which the AI system's results are received and then result-directed actions are effectuated in that external environment.[73] Such conceptions of AI and the AI lifecycle are inadequate to recognize and identify the numerous and essential points at which human decision-making impacts upon AI

---

[68] *"The Buck Stops Here" Desk sign*, HARRY S. TRUMAN LIBRARY & MUSEUM (undated), https://www.trumanlibrary.gov/education/trivia/buck-stops-here-sign (last visited Mar. 10, 2025).

[69] *See, e.g.*, *Artificial Intelligence in Society*, ORG. OF ECON. CO-OPERATION & DEV., 23 figs.1.3 & 1.4 (June 11, 2019) [hereinafter OECD] (as defined & approved by OECD AI Group of Experts (Feb. 2019)), https://www.oecd.org/en/publications/2019/06/artificial-intelligence-in-society_c0054fa1.html; Mark L. Shope, *Lawyer and Judicial Competency in the Era of Artificial Intelligence: Ethical Requirements for Documenting Datasets and Machine Learning Models*, 34 GEO. J. LEGAL ETHICS 191, 204–205 (2021) (machine learning lifecycle); Doa A. Elyounes, *"Computer Says No!": The Impact of Automation on the Discretionary Power of Public Officers*, 23 VAND. J. ENT. & TECH. L. 451, 493–95 (2021) (algorithm lifecycle).

[70] *See, e.g* NAT'L INST. OF STANDARDS. & TECH., NIST-AI.100.1 ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0),10 fig.2 (Jan. 2023), https://doi.org/10.6028/NIST.AI.100-1 [hereinafter AI RMF 1.0].

[71] *See* OECD, *supra* note 69, at 23 fig.1.4 & accompanying text. The OECD's group of experts did allow that "expert knowledge" is an input into the process of AI model-building, but it did not identify humans as the source of that knowledge.

[72] *See id.*

[73] *See id.*

systems as they are brought into being and use.[74]

This Article widens the view by presenting the idea that the most informative and therefore actionable way to conceptual the AI lifecycle is to ensure that it encompasses the entirety of AI as a human-machine enterprise. It also adds granularity to the lifecycle as necessary to enable the understanding where, within that lifecycle, AI biases come into play and the mechanisms by which they do so and interact.[75]

First, the lifecycle of AI as a human-machine enterprise proceeds along a timeline. That timeline begins at the point at which initial actions are taken to address a perceived internal or market need for an AI system. The timeline continues to run all the way through to the sunsetting of the system's use; the archival of appropriate records as to its design, development, procurement, and use, for example; and the offboarding of the system from the operational function at which it was directed and the migration, assuming the need for the function still exists, of the function and data underlying and produced by the system to a replacement system.

Second, two tracks run, largely in parallel, along this timeline. The processes captured along these tracks are generally carried out by, in one track, the user organization and, in the other track, an AI system creator internal to the user organization or, much more likely and feasible, an outside vendor that creates the system (collectively, "creator-vendor").[76] The activities along these tracks are principally independently carried out by the humans associated with the user or with the creator-vendor

---

[74] *See*, *e.g.*, Mark Haakman et al., *AI Lifecycle Models Need to Be Revised: An Exploratory Study in Fintech*, 26 EMPIRICAL SOFTWARE ENG'G 95 (July 8, 2021), https://doi.org/10.1007/s10664-021-09993-1; Thomas J. Hwang et al., *Lifecycle Regulation of Artificial Intelligence-and Machine-Learning Based Software Devices in Medicine*, 322 JAMA 2285 (Nov. 22, 2019), doi:10.1001/jama.2019.16842.

[75] More detailed and topic-specific elaborations of the lifecycle of AI as human-machine enterprise are reserved for future projects. *See*, *e.g.*, Emile Loza de Siles, *Disaggregating Artificial Intelligence Biases: A Law and Systems Engineering Approach for AI Governance and Regulation*, *in* RESEARCH HANDBOOK ON THE LAW OF ARTIFICIAL INTELLIGENCE (2d.,Woodrow Barfield & Ugo Pagallo, eds.) (Edward Elgar Pub'g, forthcoming 2024); Emile Loza de Siles *Deconstructing Artificial Intelligence Biases*, *in* GLOBAL PERSPECTIVES ON ARTIFICIAL INTELLIGENCE IMPACT ASSESSMENT (Emad Yaghmaie & Gary Marchant, eds.) (Oxford Univ. Press, forthcoming 2024).

[76] Note that there may be creators of AI systems that do not directly vend access to or use of those systems. In that case, the system creator has a contractual relationship with another organization that takes that AI solution to market. In other and now common instances, the AI creator organization makes that AI functionality available to other companies via application programming interfaces, or APIs. *See*, *e.g.*, Ben Sherry, *You Can Now Integrate ChatGPT With Your Products. Here's How*, INC. (Mar. 2, 2023), https://www.inc.com/ben-sherry/you-can-now-integrate-chatgpt-with-your-products-heres-how.html. The Article's AI lifecycle recognizes and may be adapted to address, but does not elaborate upon these potential complexities here.

organization, but at one point the activities are carried out jointly by close collaboration between these two sets of actors.[77]

1. User Track

The user track of the AI lifecycle captures all processes starting with the recognition of an internal need to create a new operational workflow or to automate, augment, or otherwise improve an existing one. Continuing along the user track, specifications for an AI system by which to meet that internal need are developed; make-versus-buy decisions are made, as applicable; if the decision is made to acquire the AI system from outside sources, *i.e.*, the much more common approach, procurement procedures are activated and carried out; the vendors for the systems design, development, implementation, and integration are evaluated, and one is selected; the contract(s) between the user and vendor organizations are negotiated and executed; and contract payment mechanisms are begun.

Although the user and creator-vendor organizations engage with one another on the prior processes along the User Track, at this point, the works becomes intensely collaborative with both organizations jointly working toward the same objective: successful deployment of the AI system into the user organization's operations.

Progressing along the user track, the system is implemented, including tailored as specific to the user organization's intended function for that system, and integrated, as needed, with other systems already in place with the organization; testing of the implementation and integrations; user training is conducted; user acceptance evaluations are completed at multiple milestones, and, if requirements met, given; standard operating procedures, or SOPs, are drafted and approved, as are continuous quality improvement, or CQI, protocols; and the system is deployed, or "goes live," into the user organization's operations; subsequent evaluations and final implementation payments are made, subject to their respective milestone requirements[78]; and ongoing licensing payments are made.

The use of the system is operationalized and continues along the user track. System audits, revalidation, and other CQI procedures,

---

[77] Note that the processes that occur along these two tracks are fact-and context-specific. The following delineation of processes is necessarily genericized for instant purposes.

[78] At this point, the period of intensive collaboration between the user and creator-vendor organization ends with the interactions between the two organizations becoming much less frequent and more periodic.

software updates, system modifications, and additional user trainings take place, or should take place, as dictated by the SOPs or otherwise as needed. For example, a machine learning model may continue to evolve as it is exposed to ever more unknown data and, potentially, feedback from its users. The model may "drift," and its computational results may become more skewed, that is, increasingly erroneous, away from the target concept.[79] The Amazon AI recruitment screening tool presents example of such drift.[80]

At some future point, the user organization's function to which the system is addressed is discontinued; vendor support for the system is discontinued; or some new system gains favor toward perceived improvements to that function. Toward the endpoint of the user track, the system is offboarded with data and other records being archived[81] subject to maintenance and requirements under the user organization's internal policies, litigation holds, government transparency legislation and regulations, or a combination thereof. Data and records may additionally be migrated to a replacement system.

2. Creator-Vendor Track

The creator-vendor track of the AI lifecycle captures all processes by which the AI system is conceptualized, specified, designed, developed, tested, introduced into the larger market, implemented and integrated, and deployed into the user's existing operations and systems, updated, and eventually, as to the user's particular instance, offboarded and its function and data potentially migrated. The drafting and refinement of business requirements starts the creator-vendor track to identify and flesh out the use case(s) that are intended to address a market or specific user need. Business and resource decisions are made as to whether the investment of people power, time, and money will yield returns that comport with the creator-vendor's acceptable risk, revenue, and profit profiles.

The next processes along the track include system requirements

---

[79] *See* Jim Holdsworth, *What Is Model Drift?*, INT'L BUS. MACH. CORP. (July 16, 2024), https://www.ibm.com/topics/model-drift; Avi Gopani, *Difference Between Concept Drift vs Data Drift in Machine Learning*, ANALYTICS INDIA MAG. (last updated Aug. 12, 2024), https://analyticsindiamag.com/concept-drift-vs-data-drift-in-machine-learning/.

[80] The Amazon machine learning tool operated in a skewed fashion due to the biases in its all-or-principally-male training data and its subsequent learning to increasingly favor male applicants over female ones. *See* notes 192–98 & accompanying text.

[81] Although presented for convenience here as a User track-terminal process, archival is ongoing throughout the period of the system's use.

and design, including as to data architectures; and data acquisition, pre-processing, annotation, and other curation and data provenance processes.[82] Where machine learning is the mode and basis the subject AI system, the creator-vendor track proceeds with the following activities[83]: the division of the bolus of data into, for machine learning, training sets and testing sets with the latter reserved for post-training evaluations; and iterative processes by which multiple models are developed or more autonomously "learned" expressing statistical correlations between features within or derived from the training data and the intended function, or target concept, of the AI system. The candidate models are evaluated, one or more selected and iteratively optimized, including through the process of applying various weights to features and adjusting other parameters, all directed toward more closely achieving the target concept. A model is selected, and one or more algorithms are created to express and enable the application of that model to unknown data, those being data other than the system's training data.

Next along the creator-vendor track, the machine learning function is fully integrated into the AI system's form factors and delivery methodology(ies).[84] For example, the creator-vendor's team works to ensure that the user interface functions well and contributes favorably to the overall user experience of the solution. Multiple rounds of testing and evaluation ensue. The AI system is productized, often being dubbed an AI "solution" in external market-facing communications. Throughout the creator-vendor track, the marketing strategy is crafted with product features being prioritized for development; the marketing message honed to position the AI solution within the desired market and to communicate its competitive advantages and marketing claims to the

---

[82] For a healthcare AI use case-focused discussion, *see generally* Hongfang Liu et al., *Artificial Intelligence Model Development and Validation*, *in* A.I. IN HEALTH CARE: THE HOPE, THE HYPE, THE PROMISE, THE PERIL 131 (Machael Matheny et al. eds., 2019), https://doi.org/10.17226/27111. For this creator-vendor discussion, the author cites to her extensive law practice experience representing Cisco, HP, and other technology companies.

[83] For this paragraph, *see* STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 651–53 (4th ed., 2022) (overview of machine learning); *id*. at 704–14 (development of machine learning systems); *id*. at 665–72 (model selection & optimization).

[84] The form factors and delivery mechanisms by which the functions of the subject AI system are presented for use vary. For example, an internally-created AI system may be an "on premise" system. A vendor-created system may be likewise, but more frequently a cloud-based model with access by authenticated users being provided via an online- or mobile device-presented user interface. These details may be important toward pinpointing where in the lifecycle AI biases arise, identifying those biases, and working to eliminate or control them. For instant purposes, such details are set aside.

markets; and pricing and distribution plans likewise being formulated. The marketing strategy is operationalized and the sales function begun.

As each user organization is identified as a potential customer, the creator-vendor track begins to align closely and then tightly dovetail with the user track. To avoid repetition, note that the creator-vendor and user collaboration begins to gain traction at the beginning stages of procurement. Assuming procurement ensues, a period of intensive collaboration ensues on through user acceptance and milestone payments, as delineated *supra* in Section I.[85] After that point, the creator-vendor track and the user track continue to interoperate in the processes described, *supra*, but much less intensively and taking on a more intermittent and "maintenance"-focused character.

For simplicity, the Article has cast its discussion in this subsection as involving two actor organizations. Easily, more organizations could be at work throughout the lifecycle of AI as a human-machine enterprise. Each of those organizations, however, is comprised of and acts through and only through people. Each of these human beings make decisions, take actions, do both, or contribute to same. Those many decisions and actions dictate the entirety of the AI system and its functioning or malfunctioning, including as to errors that result from AI biases. The more granular and comprehensive AI lifecycle laid out here renders much more visible the many points of human agency within AI as a human-machine enterprise. In addition, it brings to sight the many points during that lifecycle at which AI biases may be injected or arise and, so, makes actionable progress possible.

III. AI BIASES, A TAXONOMY AND BEGINNING COMPENDIUM

To facilitate actionable intelligence about AI biases so that humans may act with intentionality and responsibility, the two sets of action are important. First, each of the many AI biases must be identified, described, and its operations and impacts understood. Second, the problem space must be organized, and the biases categorized into a taxonomy for action. The detailed AI lifecycle presented in the prior Section connects with both of these action sets, providing an essential organizational foundation toward the second action set and thereby enabling the first.

Toward the first action set, the author has developed a beginning

[85] *See* note 78 & accompanying text.

compendium of fifty AI biases.[86] Space here precludes the compendium's full presentation. Toward legitimating the AI biases taxonomy and discussions that follow, an overview of the author's work in creating the compendium is important. The author synthesized the AI biases compendium through novel translational and interpretive efforts. She first read extensive technical literatures about AI biases and then translated the findings into language more accessible to readers in the law and others generally outside the scientific and technical disciplines. Next, she read and incorporated understandings from bias literatures of the statistical, economic, and epidemiological disciplines.[87] Here, an extensive search of federal court decisions, federal regulations, and legal scholarship was carried out to identify sources discussing any of the fifty AI biases, including as known by other terms. Finally, the author analyzed, interpreted, and synthesized these understandings of AI biases to create the compendium and explore and continue to develop specific illustrative AI bias use cases.

As to the second, this Article organizes the AI biases within a taxonomy of six AI bias categories: (1) cognitive biases; (2) societal, institutional, and other cohort-held biases (collectively in this paragraph, "societal biases"); (3) data biases; (4) learning biases; (5) model biases; and (6) use biases.

Throughout the AI lifecycle and within this Article's AI biases taxonomy, two categories of AI biases are overarching. Cognitive biases, those held by individual humans, and societal biases, those held by groups, are global and pervasive. These two types of "umbrella" biases have or have the potential to have a global impact upon the AI human-machine enterprise. They have their own impacts and have impacts, that is, some causal or contributory link to the other four categories of AI biases.

Those four categories, *i.e.*, data, learning, model, and use biases, collectively map to four or more spans within the AI lifecycle. AI biases along these spans may relate to human engagement in the user track, the creator-vendor track, or both and in varying proportions. For example, the learning category of AI biases may be primarily concentrated, although not exclusively, in the lifecycle span that covers model learning and optimization and specifically along the creator-vendor lifecycle track

---

[86] *See* Emile Loza de Siles, Artificial Intelligence Biases (2022-present) (unpublished manuscript) (on file with author).

[87] This step was important, in part, to clarify the ununified and highly variable nomenclature used in the discussion to AI biases in the technical literatures and ground those usages within the often more foundational understandings of biases generally.

within that span. On the other hand, the data category of AI biases, despite being frequently treated as presenting issues pertaining only to training data and AI developers, may originate with the user organization. Even AI biases that may seem to originate and operate along narrow spans of the AI lifecycle biases and along just one of the two lifecycle tracks, however, may impact across the other track and other parts or the entirety of the lifecycle. Most importantly, even narrowly operating AI biases may produce broad, significant, and long-tailed first-order impacts upon the people who are exposed to them and additional second and third-order impacts beyond that.

In the following subsections, the Article briefly describes each of the two global and four more discretely operating categories of AI biases. It identifies the types of AI biases that generally fit within each category.[88] Per category, it discusses one exemplar bias in some detail, providing brief backgrounders where needed to orient that discussion. It closes each subsection by presenting an illustrative use case for that exemplar AI bias.[89]

A. Cognitive Biases

A **cognitive bias** is "the way a particular person understands events, facts, and other people, which is based on their own particular set of beliefs and experiences and may not be reasonable or accurate."[90] More fully,

> [c]ognitive biases are systematic cognitive dispositions or inclinations in human thinking and reasoning that often do not comply with the tenets of logic, probability reasoning, and plausibility. These intuitive and subconscious

[88] Although some biases may present in multiple AI bias categories, this Article intentionally assigns AI biases to a single category for clarity and simplicity with elaborations being reserved for a future work.

[89] The choice of exemplars generally rests upon the facility with which the subject bias may be translated or interpreted, as highlighted in the text; the availability of reliable and relatively complete information by which to illustrate that bias with an AI use case; and the author's view as to the relative importance of understanding the exemplar bias.

[90] *Cognitive bias*, CAMBRIDGE ADVANCED LEARNER'S DICTIONARY & THESAURUS DICTIONARY (undated), https://dictionary.cambridge.org/us/dictionary/english/cognitive-bias (last visited Aug. 9, 2024). *See* NIST REP'T, *supra* note 30, at 49; Margaret A. Berger, *The Admissibility of Expert Testimony*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 29 (3d ed. 2011) [hereinafter SCIENTIFIC EVIDENCE]; Paul C. Giannelli et al., *Reference Guide on Forensic Identification Expertise*, *in* REFERENCE MANUAL ON SCI. EVIDENCE 79 (3d ed. 2011); John B. Wong et al., *Reference Guide on Medical Testimony*, *in* REFERENCE MANUAL ON SCI. EVIDENCE 706 (3d ed. 2011).

> tendencies are at the basis of human judgment, decision making, and the resulting behavior. Psychological frameworks consider biases as resulting from the use of (inappropriate) cognitive heuristics that people apply to deal with data-limitations, from information processing limitations, or from a lack of expertise.[91]

These two definitions of cognitive bias lean toward the concept of implicit bias. Implicit bias is a powerfully important cognitive bias and has been the subject of extensive scholarly scrutiny.[92] Implied bias, however, is only one type of cognitive bias, and even it operates across multiple topical domains.[93]

Reports vary, but there may be at least 188 types of cognitive biases.[94] Numerous scholars in law, behavioral economics, and other disciplines have considered the cognitive biases within legal and intersecting domains.[95] This Article identifies at least ten cognitive biases as being AI-relevant.[96] Alphabetically ordered, those ten AI biases are

---

[91] J.E. (Hans) Korteling & Alexander Toet, *Cognitive Biases*, *in* ENCYCLOPEDIA OF BEHAVIORAL NEUROSCIENCE 610, 610 (Sergio Della Sala, ed., 2d ed., 2022), https://doi.org/10.1016/B978-0-12-809324-5.24105-9.

[92] *See generally*, *e.g.*, Charles R. Lawrence III, *Unconscious Racism Revisited: Reflections on the Impact and Origins of "The Id, the Ego, and Equal Protection"*, 40 CONN. L. REV. 931 (2008); Justin D. Levinson et al., *Judging Implicit Bias: A National Empirical Study of Judicial Stereotypes*, 69 FLA. L. REV. 63 (2017).

[93] *See* PROJECT IMPLICIT (undated), https://implicit.harvard.edu/implicit/takeatest.html (last visited Nov. 18, 2024) (offering implicit association tests across various bias foci, *e.g.*, gender-career, transgender-cisgender, age, disability, skin color).

[94] Jeff Desjardins, *Every Single Cognitive Bias in One Infographic*, VISUAL CAPITALIST (Aug. 26, 2021), https://www.visualcapitalist.com/every-single-cognitive-bias/. Hellström and his co-authors point to a crowd-sourced reference indicating that there are at least 190 types of cognitive bias. *See* Hellström et al., *supra* note 24, at 3. These co-authors assert that only a few cognitive biases relate directly to machine learning, but they do not offer any basis for that assertion. *See id.*

[95] *See, e.g.*, Ryan Calo, *Digital Market Manipulation*, 82 GEO. L. REV. 995, 1001-02 (2014) (observing few government enforcements against exploitation of consumers' cognitive biases); Caleb S. Fuller, *Is the Market for Digital Privacy a Failure?,* 180 PUB. CHOICE 353, 356 (2019), https://doi.org/10.1007/s11127-019-00642-2 (examining survey data and other work to conclude that personalization and other digital information-collection practices exploit consumers' cognitive biases to extract increasingly quanta of data from and about them, despite their demonstrated dislike of those practices). *See generally* Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161 (1995).

[96] There are almost assuredly more than ten cognitive biases that are AI-relevant.

anchoring bias [97]; automation bias [98]; availability bias [99]; class membership bias; confirmation bias[100]; two biases that are here coined as "face-saving" bias [101] and "incentivized viewpoint" bias [102]; the aforementioned implicit bias [103]; interpretation bias [104]; and loss of situational awareness bias.[105]

As the chosen exemplar, class membership bias is a cognitive bias in which certain inferences are made about an individual based upon their membership within a group, or class.[106] Racial bias, gender bias, and anti-trans bias, for example, are types of class membership bias. Class membership bias may enter into and impact decision-making through heuristics that categorize, for example, persons of color as more prone to criminality.[107] Such class membership bias may then impact pretrial detention or sentencing decisions based upon the perceived risks of recidivism or violence that result in harsher outcomes for persons of color than for non-Hispanic Caucasian persons similarly situated within such criminal law contexts.[108]

The injection of class membership bias into the AI human-machine enterprise, however, may occur at points other than at which humans consume and act upon the AI system's output. For instance, class membership bias also may be injected into the computational underpinnings of AI systems during their development and post-deployment maintenance and updating. Homogeneity within AI design and development teams likely increases the risk that class membership

---

[97] *See* NIST REP'T, *supra* note 30, at 49; Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 75 & n.109 (2019).
[98] *See* EU AI Act, *supra* note 15, at Art. 14(4)(b); Loza de Siles, *AI, on the Law of the Elephant*, *supra* note 24, at 1406–08; Danielle Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1262, 1271 (2008); *Heuristic*, OXFORD ENG. DICTIONARY (3d. ed. 2014) (Entry B2).
[99] *See* Daryl Lim, *Predictive Analytics*, 51 LOY. U. CHI. L. J. 161, 216–19 (2019); William Magnuson, *Artificial Financial Intelligence*, 10 HARV. BUS. L. REV. 337, 361–62 (2020); NIST REP'T, *supra* note 30, at 49.
[100] *See* Lim, *supra* note 99, at 216–19; NIST REP'T, *supra* note 30, at 50.
[101] *Accord* NIST REP'T, *supra* note 30, at 50 (Dunning-Kruger effect).
[102] *Accord id.* at 51 (funding bias).
[103] *See id.* at 52.
[104] *See id.*; Kevin M.K. Fodouop, Note, *The Road to Optimal Safety: Crash-Adaptive Regulation of Autonomous Vehicles at the National Highway Traffic Safety Administration*, 98 N.Y.U. L. REV. 1358, 1381 & n.106 (2023).
[105] *See* NIST REP'T, *supra* note 30, at 52.
[106] *See id.* at 23, 50 (dubbed "ecological fallacy").
[107] *See* Mayson, *supra* note 44, at 2277–79.
[108] *See generally, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 AM. CRIM. L. REV. 1553 (2019).

bias to the extent that the team members, their groupthink, or both reflect stereotypical views associated with classes encompassed within the system under development.[109] Such stereotypical views and their corresponding class membership biases may find their ways into the selection and weighting of features within the datasets with which AI system models are found, trained, selected, and optimized.[110] For example, class membership bias may be injected in development decisions as to the weightings of features used to predict an individual's performance in collegiate education, in part, based upon their race.[111]

The NFL's now-discredited and barred use of race norming exemplifies the injection of class membership bias into a computational system and, specifically, its algorithmic design.[112] Race norming was invoked during the computation of former NFL players' degree of cognitive impairment post-concussive injuries by comparison to their purportedly pre-injury capacities, which were systematically lowered based upon their non-White racial and ethnicity class memberships.[113]

B. Societal and Other Cohort Biases

Societal and other cohort biases are those that are or may be prevalent and broadly reflected with a particular societal group, cultural group, institution,[114] company, or other cohort of people. These cohort biases differ from cognitive biases because they reflect groupthink and collective perspectives shared by the group, whereas cognitive biases are those reflected in the thinking and decision-making of a single individual. There are at least six AI biases with this category of societal and other

---

[109] *See* Rifat Ara Shams et al., *AI and the Quest for Diversity and Inclusion: A Systematic Literature Review*, 13 AI & ETHICS (2023), https://doi.org/10.1007/s43681-023-00362-w.

[110] *See generally, e.g.*, note 108.

[111] *See* Todd Feathers, *Major Universities Are Using Race as a "High Impact Predictor" of Student Success*, THE MARKUP (Mar. 2, 2021), https://themarkup.org/news/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success, cited in NIST REP'T, *supra* note 30, at 23.

[112] *See* Dave Zirin, *So What the Hell Is Race Norming?*, THE NATION (Mar. 12, 2021), https://www.thenation.com/article/society/race-norming-nfl-concussions/. Race norming is a pre-analytical statistical method by which the data are adjusted for the intended purpose of reducing or removing systemic biases in those data. *See id.*

[113] Although race norming has often been discussed as presenting a Black-White dichotomy, the practice of race norming has long been properly understood as far broader in its reach. *See* Michael A. Olivas, *Legal Norms in Law School Admissions: An Essay on Parallel Universes*, 42 J. LEGAL EDUC. 103 (1992).

[114] *See* Ifeoma Ajunwa, *The Paradox of Automation As Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1697–98 (2020); NIST REP'T, *supra* note 30, at 52.

cohort biases, as follows: historical bias[115]; persistent systemic inequality bias[116]; co-occurrence bias[117]; epistemological bias; framing bias; and language bias.[118]

As a preliminary matter, historical bias is conceptually a broader category of bias than is persistent systemic inequality bias. As an example of the former, consider that the cohort of those of driving age in the United States has long been biased in favor of the automobile over mass transit. This bias has little to nothing to do with persistent systemic inequality bias. [119] Much has been written about AI and persistent systemic inequality bias, [120] a vitally important topic on which scholarship, law and policy development must and does, here, continue. Toward a novel contribution, however, this Article focuses its exemplar on another societal or cohort AI bias: framing bias.

Framing biases, together with epistemological and language biases, are those biases reflected in text data and the context within which the text appears.[121] To focus on a singular example, framing bias occurs

---

[115] *See* Hellström et al., *supra* note 24, at 2, 7, Fig.1; *accord also* NIST REP'T, *supra* note 30, at 52. The NIST report mentions "historical bias," but in the context of structural inequity and eschewing any definition, examination, or elaboration. *See* NIST REP'T, *supra* note 30, at 51.

[116] *See, e.g.*, Brown v. Board of Educ. of Topeka, 347 U.S. 483 (1954); Kenneth B. Clark & Mamie P. Clark, *Emotional Factors in Racial Identification and Preference in Negro Children*, *in* READINGS IN SOCIAL PSYCHOLOGY 169 (Eleanor E. Maccoley, Theodore M. Newcomb & Eugene L. Hartley, eds., 1947); Erin Blakemore, *How Dolls Helped Win Brown v. Board of Education*, HISTORY, https://www.history.com/news/brown-v-board-of-education-doll-experiment (last updated Sept. 29, 2023).

[117] *See* Cheongwoong Kang & Jaesik Choi, *Impact of Co-occurrence on Factual Knowledge of Large Language Models*, ARXIV (Oct. 12, 2023), https://arxiv.org/abs/2310.08256.

[118] *See* Hellström et al., *supra* note 24, at 1–2, 7, Fig.1; Mackenzie Pike, *"Can You Repeat That?": Why AI Speech Discrimination Should Decelerate the Use of Automated Speech Recognition as a Medical Record Tool*, 31 ANNALS HEALTH L. ADVANCE DIRECTIVE 193, 197–98 (2021).

[119] Other writings treat these two biases as synonymous, however. *See, e.g.*, Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. 1, 33–36, 51 (2021). *See generally* Sandra Wachter et al., *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law*, 123 W. VA. L. REV. 735 (2021).

[120] *See generally, e.g.*, Emile Loza de Siles, *Artificial Intelligence Bias and Discrimination: Will We Pull the Arc of the Moral Universe Toward Justice?*, 8 J. INT'L & COMP. L. 513 (2021); Emile Loza de Siles, *Soft Law for Unbiased and Non-Discriminatory Artificial Intelligence*, 40 IEEE TECH. & SOC. MAG., SPEC. ISSUE ON SOFT L. GOVERNANCE OF A.I. 77 (Dec. 2021); Virginia Eubanks, DIGITIZING THE CARCERAL STATE: AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016); Sahar Takshi, *Unexpected Inequality: Disparate-Impact from Artificial Intelligence in Healthcare Decisions*, 34 J.L. & HEALTH 215 (2021).

[121] *See* Hellström et al., *supra* note 24, at 2–3.

when the text expresses an opinion about a particular topic.[122] To examine this bias, Canadian researchers Svetlana Kiritchenko and Saif Mohammad undertook analyses of 219 natural language processing ("NLP") AI systems against their Equity Evaluation Corpus ("EEC") database of almost 9000 carefully selected English language sentences.[123] They controlled all variables to isolate and test the sentiment predictions of those NLP systems for gender and a variety of racial biases.[124]

Seventy-five percent (75%) of those systems demonstrated gender-based framing biases in predicting sentiments reflecting, for instance, stereotypes that hold women to be more emotional and men less so.[125] Race-based framing biases were even more prevalent in the systems.[126] For example, an even larger proportion of the studied systems comparatively predicted the intensity of the negative emotions of fear, anger, and sadness as greater where the EEC sentences were connected with "African American" indicators, but of the positive emotion of joy as greater where the sentences were connected with "European American" indicators.[127] Framing bias occurs well beyond racial and gender classes, however. For instance, sentiment predictors have rated text discussing "disability" as negative while AI-powered content moderators designate such text as "toxic."[128]

Like cognitive biases, social and cohort biases may affect the AI human-machine enterprise, arising at any and all phases of the AI lifecycle. Next, the Article turns to AI biases that occur in more distinctive phases of that lifecycle and impact upon still others.

C. Data Biases

"Data-centric AI" is a relatively recently emerged philosophy and approach to artificial intelligence championed and practiced by some of the leading minds in the field.[129] Among its other attributes, data-centric

---

[122] *See id.*
[123] Svetlana Kiritchenko & Saif M. Mohammad, *Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems*, *in* PROC. OF 7TH JOINT CONF. ON LEXICAL & COMPUTATIONAL SEMANTICS 1–2 (May 11, 2018), https://arxiv.org/abs/1805.04508.
[124] *See id.*
[125] *Id.* at 5–9.
[126] *See id.* at 9.
[127] *See id.* at 8–9.
[128] MEREDITH WHITTAKER ET AL., AI NOW INST., DISABILITY, AI, AND BIAS 8 (2019), https://ainowinstitute.org/publication/disabilitybiasai-2019.
[129] Eliza Strickland, *Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big*, 59 IEEE SPECTRUM 23, 24 (2022), doi: 10.1109/MSPEC.2022.9754503; *see generally id.*

AI focuses upon the quality, character, and availability of data as prime to the machine learning, modeling, and other AI development practices that follow, and data biases are therefore of great importance.

Data biases are those biases that exist or are injected into the subject data. "Garbage in, garbage out," expresses the intuition behind data biases.[130] At base, the problem with data biases is that AI models developed, trained, tested, validated, and optimized upon and using biased datasets cannot be, in any objective sense, accurately and reliably generalize predictions or recommendations when applied to unknown data obtained from the broader world and ingested into the consequently fundamentally-erroneous AI system. [131] If not recognized and then eliminated, if possible, or otherwise controlled for, AI biases within data contribute to and exacerbate other biases downstream in the AI lifecycle. Biases may not be recognized until later learning and modeling phases of the AI lifecycle, but their origin stories rest within the earlier data phase.

[130] Simon Jelley, *Garbage In, Garbage Out: The Role of Data Management in Effective AI*, FORBES (Nov. 16, 2023), https://www.forbes.com/sites/forbesbusinesscouncil/2023/11/16/garbage-in-garbage-out-the-role-of-data-management-in-effective-ai/?sh=7e2646fbdbb0.
[131] *Accord*, *e.g.*, Hellström et al., *supra* note 24, at 3–4, 7, Fig.1; NIST REP'T, *supra* note 30, at 15.

At least fourteen AI biases fall within the category of data biases. Those are activity bias[132]; annotation bias, discussed *infra*; class imbalance bias[133]; coverage bias[134]; image dataset bias[135]; information bias,[136] of which there are several subtypes, including measurement bias, observational bias, and misclassification bias[137]; negative set bias[138]; sampling bias, also called selection bias[139]; representation bias[140]; representativeness bias[141]; and specification bias.[142] Here, the Article discusses annotation bias as an exemplar AI data bias after providing a brief orientation to data labeling, or annotation.

---

[132] *See* Ricardo Baeza-Yates, *Bias on the Web*, 61 COMM'S OF THE ACM 54, 56–57 (2018), http://dx.doi.org/10.1145/3209581; Ricardo Baeza-Yates & Diego Saez-Trumper, *Wisdom of the Crowd or Wisdom of a Few? An Analysis of Users' Content Generation*, *in* PROC'S 26TH ACM CONF. ON HYPERTEXT & SOC. MEDIA 69–74 (2015), https://dl.acm.org/doi/10.1145/2700171.2791056; Katyal, *supra* note 97, at 71. *See also* Gillian K. Hadfield, *Bias in the Evolution of Legal Rules*, 80 GEO. L. J. 583, 597 (1992).

[133] *See* Erdal Tasci et al., *Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets*, 14 CANCERS 12 (2022). *See* Method for Detecting and Mitigating Bias and Weaknesses in Artificial Intelligence Training Data and Models, U.S. Patent No. 11,256,989, col. 20 (issued Feb. 22, 2022) [hereinafter, '989 Patent].

[134] *See* Hellström et al., *supra* note 24, at 3, 7; Shuo Wang et al., *On the Language Coverage Bias for Neural Machine Translation*, ARXIV (June 7, 2021), https://arxiv.org/pdf/2106.03297.pdf; Mary H. Mulry, *Coverage Error*, *in* ENC. OF SURVEY RES. METHODS 162, 162 (Paul J. Lavrakas, ed., 2008), https://dx.doi.org/10.4135/9781412963947.n115. *See*, *e.g.*, Anja Mohorko, Edith de Leeuw & Joop Hox, *Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time*, 29 J. OFF. STAT. 609 (2013), https://doi.org/10.2478/jos-2013-0042.

[135] *See* Sonia K. Katyal & Jessica Y. Jung, *The Gender Panopticon: AI, Gender, and Design Justice*, 68 UCLA L. REV. 692, 709–18 (2021).

[136] *See* Dov Greenbaum, *Direct Digital Engagement of Patients and Democratizing Health Care*, 32 SANTA CLARA HIGH TECH. L. J. 93, 135–36 (2016).

[137] This Article adopts the hierarchy expressed in the accompanying text and attempts to thereby remedy inconsistencies across various sources' labelling of the subject biases. *See* SCIENTIFIC EVIDENCE, *supra* note 90, at 583; *id.* at 585 ("Information bias is a result of inaccurate information about either the disease or the exposure status of the study participants or a result of confounding."); *id.* at 624 (identifying information bias synonymous to observational bias). Other sources treat information bias as synonymous to measurement bias. *Compare*, *e.g.*, 2014 DICTIONARY, *supra* note 24, at 149 (second definition of information bias) *with id.* at 180 (measurement bias). Others categorize only biases arising from certain types of measurement errors as being information bias. *See* Gaël P. Hammer et al., *Avoiding Bias in Observational Studies*, 106 DEUTSCHES ÄRZTEBLATT INT'L 664, 665 (2009), doi: 10.3238/arztebl.2009.0664.

[138] *See* Antonio Torralba & Alexei A. Efros, *Unbiased Look at Dataset Bias*, *in* PROC. OF 24TH IEEE CONF. ON COMPUTER VISION & PATTERN RECOGNITION 1521, 1525–26, 1528 (2011), doi: 10.1109/CVPR.2011.5995347; Tatiana Tommasi, et al., *A Deeper Look at Dataset Bias*, ARXIV 2–3 (2015), https://arxiv.org/ftp/arxiv/papers/1505/1505.01257.pdf.

[139] *See* Hellström et al., *supra* note 24, at 3–4; NIST REP'T, *supra* note 30, at 15.

Sampling bias is the error introduced when individuals within a population are not equally likely to have been included within the sample of that population and, consequently, the sample does not accurately represent the entire population. *See* Julia Simkas, *Sampling Bias: Types, Examples & How to Avoid It*, SIMPLY PSYCH. (last updated July 31, 2023), https://www.simplypsychology.org/sampling-bias-types-examples-how-to-avoid-it.html.

The 2020 United States Census, for example, exhibited significant sampling bias. Its sampling bias as to Hispanics tripled in comparison to that of the 2010 census, undercounting this growing population by more than five percent. Sampling biases in the 2020 Census also resulted in undercounting across numerous other populations of color, but the overcounting of non-Hispanic Caucasians and Asians. *See US Census Undercounted Minorities in 2020, New Data Shows*, AL JAZEERA (Mar. 10, 2022) [hereinafter *US Census Errors*], https://www.aljazeera.com/news/2022/3/10/us-census-undercounted-minorities-2020-new-data-shows. Among other negative impacts, the perhaps most unjust and impactful consequences of these Census sampling biases is that they undermine the bedrock of equal representation in Congress and equal protection under the U.S. Constitution.

The Federal Judicial Center of the National Research Council and other authorities use the term "**selection bias**" to discuss this concept of bias created by the manner in which data are sampled. *See* SCIENTIFIC EVIDENCE, *supra* note 90, at 224–25 (discussing selection bias in pedestrian surveys where only a portion of pedestrians are approached for survey participation). *See also*, *e.g.*, Drew DeSilver, *Just How Does the General Election Exit Poll Work, Anyway?*, PEW RES. CTR. (Nov. 2, 2016), https://www.pewresearch.org/fact-tank/2016/11/02/just-how-does-the-general-election-exit-poll-work-anyway/ (discussing surveys of voters exiting polls on election day) (noting augmentation by phone surveys, given significant election participation by mail-in balloting, rather than by in-person voting at polls).

[140] For more on data representation, *see Different Forms of Data Representation in Today's World*, GEEKS FOR GEEKS (last updated Sept. 14, 2021), https://www.geeksforgeeks.org/different-forms-of-data-representation-in-todays-world/; ROHAN CHOPRA ET AL., DATA SCIENCE WITH PYTHON (2019), *accessible copy of cited material available at* https://subscription.packtpub.com/book/data/9781838552862/1/ch01lvl1sec04/data-representation (last visited Jan. 31, 2025); C. M. Sperberg-McQueen & David Dubin, *Data Representation*, § 1 DIGITAL HUMANITIES DATA CREATION (undated), https://archive.mith.umd.edu/dhcuration-guide/guide.dhcuration.org/index.html%3Fp=63.html (last visited Jan. 31, 2025); NIST REP'T, *supra* note 30, at 18 ("inexact representation").

[141] *See* Tufekci, Zeynep, *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*, *in* ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (2014), https://arxiv.org/pdf/1403.7400. Some behavioral finance sources consider representativeness bias to be a type of cognitive bias. *See generally*, *e.g.*, Charles Chang et al., *A Test of the Representativeness Bias Effect on Stock Prices: A Study of Super Bowl Commercial Likeability*, 103 ECON. LETTERS 49 (2009), https://doi.org/10.1016/j.econlet.2009.01.018. *See also* Thomas F. Cotter, *Patent Damages Heuristics*, 25 TEX. INTELL. PROP. L.J. 159, 166 (2018).

[142] *See* Hellström et al., *supra* note 24, at 3, 7, Fig.1; James J. Heckman, *Sample Selection Bias as a Specification Error*, 47 ECONOMETRICA 153, 153 (1979) ("This paper discusses the bias that results from using nonrandomly selected samples to estimate behavioral relationships as an ordinary specification bias that arises because of a missing data problem."), *quoted in* J.J. Prescott et al., *Understanding Noncompetition Agreements: The 2014 Noncompete Survey Project*, 2016 MICH. ST. L. REV. 369, 464 n.300 (2016).

1. A brief backgrounder on annotation

Annotation refers to data annotation, also known as data labeling. Data annotation generally occurs as part of the data curation, training, and testing processes or a combination thereof within the AI lifecycle.[143] **Annotation** is the process of determining and associating tags, captions, and other labels[144] with data so that they may be subsequently input into the subject AI system and so that the system can "understand" them.[145] This means that the data must made structured or otherwise fit, through annotations, to be ingested in the AI system and thus used within its computations.[146] Human beings,[147] automated systems,[148] or combinations of both[149] create these annotations.

---

[143] These processes may be integrated and not sequential or iterative or both, and persons engaged in annotation operations may be employed within the developing organization or performing the work under one or more contract.

[144] This Article uses the term "label" to refer to the words or tags that are associated with input data. Some other sources use the word "label" to also refer to the output of any given machine learning system, such as the system's computed prediction as to what kind of animal is depicted in an image or a risk score, for instance. *See Framing an ML Problem*, GOOGLE DEVELOPERS, https://developers.google.com/machine-learning/problem-framing/ml-framing (last updated Oct. 27, 2023).

[145] Like human understanding, "machine understanding" of data is hyperbole. *See* IAN H. WITTEN & EIBE FRANK, DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES WITH JAVA IMPLEMENTATIONS 2 (2000) ("We would all testify to the growing gap between the *generation* of data and our *understanding* of it. As the volume of data increases, inexorably, the proportion of it that people 'understand' decreases, alarmingly.") (emphasis in original).

[146] Data sourced and curated for use in machine learning and other types of AI may be generally categorized as structured or unstructured. *See* Loza de Siles, *supra* note 24, at 1446–48. Structured data rest in orderly fashion in spreadsheets and other configurations that render the data easily and directly accessible within the subject AI lifecycle processes and corresponding AI system. Structured data constitute the minority of data.

By contrast, unstructured data comprise the majority, ranging, for example, from "textese" language and emoticons from instant messages; videos, photographs, memes, and social media avatars to the obscure and arcane terms of a 101-year-old, but still operational, banking contract. *See* Seema Phekoo, Managing Counsel, BNY Mellon, Artificial Intelligence, Analytics, and Today's Future-Forward Law Practice, Guest Lecture in Author's Artificial Intelligence and Social Justice Course, 18:12–18:14 (Nov. 16, 2021) (video on file with author). Annotation and other data processing steps are required to render unstructured data capable of use within the AI lifecycle and system.

[147] As examples, data annotation by humans may be crowdsourced from among globally distributed individuals, *see*, *e.g.*, AMAZON MECHANICAL TURK (undated), https://www.mturk.com/ (last visited Mar. 10, 2025) [hereinafter MTurk]; outsourced and technology-augmented project teams, *see*, *e.g.*, *Amazon SageMaker Ground Truth Features: Data Labeling*, AMAZON WEB SERVS. (undated), https://aws.amazon.com/sagemaker/groundtruth/features/?nc=sn&loc=2 (last

In addition to labeling AI input data, annotators label output data.[150] In machine learning and, specifically, supervised learning, input and output data are paired so that the AI system may find models that attempt to correlate those pairs. The correlations, however, are between annotations associated with the input data with those associated with the output data.[151] After process to evaluate candidate models and then select and optimize the selected model and the AI system rolled out for use, the underlying fact remains that these data "in the wild" are compared against these annotation-enabled correlations between AI input and output data.[152]

2. Annotation bias

To greater or lesser degrees, annotations create "noise" within the data.[153] The presence of such noise means that the data, as annotated, are incomplete or otherwise differ from the original underlying data.[154] This

---

visited Mar. 10, 2025); or in-house teams, *see* Loza de Siles, *supra* note 24, at 1432 n.151.

For annotating natural language processing, or NLP, datasets, crowdsourcing has become the predominant method. *See* Mor Geva et al., *Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets*, *in* PROC. OF THE 2019 CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING & 9TH INT'L JOINT CONF. ON NAT. LANGUAGE PROCESSING 1161, 1161 (Nov. 2019), https://aclanthology.org/D19-1107.pdf.

[148] *See* Kyle Wiggins, *MIT Study Finds 'Systematic' Labeling Errors in Popular AI Benchmark Datasets*, VENTUREBEAT (Mar. 28, 2021), https://venturebeat.com/business/mit-study-finds-systematic-labeling-errors-in-popular-ai-benchmark-datasets/.

[149] *See*, *e.g.*, Kyle Johnson, Head of Compliance & Regulatory Services, BNY Mellon, Artificial Intelligence, Analytics, and Today's Future-Forward Law Practice, Guest Lecture in author's Artificial Intelligence and Social Justice Course, 18:14–18:18 (Nov. 16, 2021) (video on file with author).

[150] *See* Hellström et al., *supra* note 24, at 4.

[151] *See generally* Geva et al., *supra* note 147, at 1161.

[152] "In the wild" is a commonly used expression meaning that the subject system has been deployed into the market and operations and is being applied to live instances of real data in real time to predict a given output, *e.g.*, a predictive risk classification score as to whether an individual should be approved or rejected for the extension of credit. *See*, *e.g.*, Hellström et al., *supra* note 24, at 4.

[153] *See* Ishan Misra et al., *Seeing Through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels*, *in* PROC. 2016 IEEE CONF. ON COMP. VISION & PATTERN RECOGNITION 2930, 2930–32 (Jun. 2016), http://ieeexplore.ieee.org/document/7780689/. This noise may be more specifically "model labeling noise." *Id.* at 2931.

[154] Some researchers studying this phenomenon and working to design machine learning systems by which to computationally annotate data have dubbed such

deviation between the ground truth of the information within those original data and the information within the data as annotated comprises **annotation bias**.[155]

Considering image data, for instance, these annotations constitute descriptors of the visually rich contents of these images.[156] The annotations to be captured may vary depending upon the aims of the annotation project at hand. In the work to create the Visual Genome dataset, for example, densely annotated image data may be desired to document the objects shown within images, attributes of those objects, relationships between the objects, descriptions reflecting the regions depicted therein.

As noted above, annotation bias may arise in several ways. For one exemplar, annotators, human or otherwise, may fail to include image content within the scope of their annotations, the incompleteness of the annotations introduces bias, and that bias may be impactful when the annotated data are subsequently used by AI systems. For instance, annotations of an image containing bananas may fail to include a tag of "yellow." Annotators may consider the color yellow to be irrelevant because, for instance, the cloned yellow Cavendish banana predominates

---

failures to be "human-centric" annotation, which results from human reporting bias. *Id.* at 2931.

Identifying annotation bias as "human-centric" mischaracterizes the problem on at least three counts, however. First, it is more accurate is to ascribe such bias to a "culture-centric annotation" because the error cannot be generalized to all human annotators and especially not where it is culturally and geographically diverse. *See*, *e.g.*, Jeanne Whalen & Yuan Wang, *Hottest Job in China's Hinterlands: Teaching AI to Tell a Truck from a Turtle*, WASH. POST (Sept. 26, 2019), https://www.washingtonpost.com/business/2019/09/26/hottest-job-chinas-hinterlands-teaching-ai-tell-truck-turtle/.

As a closely related second, the atomization of annotation operations across large globally distributed networks of annotation workers with language, literacy, and cultural differences may inject further biases into the human-machine enterprise of AI. *Accord*, *e.g.*, Whalen & Wang, *supra* note 154 ("We can't understand the text . . . . We just draw a frame around the image to crop the text — that's all we're responsible for."). *See generally*, *e.g.*, MTURK, *supra* note 147 (discussing atomization of annotation into "microtasks," such as entering five tags per images, for online execution by distributed workers). Third, the misnomer suggests, incorrectly, that annotation-by-AI is immune from biases that its human creators may possess.

[155] For convenience, this Article generally refers to annotation bias in the singular. Note, however, that there is more than one type of annotation bias. Note, also, that some sources use the terms "category bias" and "label bias" to ascribe the meaning of annotation bias as used in this Article. *See*, *e.g.*, Torralba & Efros, *supra* note 138, at 1525; '989 Patent, *supra* note 133, at 22, col. 19 (synonymizing category bias & label bias). Finally, the author does not advise reliance upon the NIST Report's discussion of annotation bias.

[156] These descriptors may take a variety of grammatical forms, including adjectives, nouns, or gerunds, phrases, or even whole sentences, *See*, *e.g.*, Geva et al., *supra* note 147, at 1162.

Western markets and as the world's largest fruit crop.[157] By contrast, banana eaters from non-majority cultures and in other parts of the world enjoy fruits, from among its more than 1,000 types, that have red, pink, purple, or even black skins.[158]

"Banana bias" may seem trite but decidedly is not. A recent Massachusetts Institute of Technology ("MIT") study evaluated ten of the most highly-cited benchmark datasets used to test machine learning systems [159] and previously assumed to be correct "gold standard" reference standards for that purpose. [160] MIT researchers Curtis Northcutt, Anish Athalye, and Jonas Mueller found that all of these datasets are plagued with prevalent and systematic labelling errors.[161] On average, the contents of these test datasets had a labelling error rate of at least 3.3%.[162] The individual error rates of some of these datasets were much higher. [163] For example, the ImageNet validation set contained 2,916 labeling errors, that is, in 6% of the entire image dataset, which is categorized into 1,000 classes.[164] The MIT study estimated more than 5 million labelling errors, representing more than 10% of the "Quick, Draw!" dataset's 1 billion images.[165]

Annotation bias is indicative of larger bias problems that arise through data annotation across the great and increasing numbers of AI applications that use unstructured data. These bias problems, in turn, can result in errors in the models "learned" by AI systems and onward through the AI lifecycle into the live use of the system. Properly training human or machine annotators is a non-trivial matter but has been shown

---

[157] *See Bananas – The Most Popular Fruit in the World*, ALLFRESCH (undated), https://www.allfreschgroup.com/bananas-the-most-popular-fruit-in-the-world/ (last visited July 27, 2024).
[158] *Id.*; *see*, *e.g.*, *Red Bananas*, SPECIALTY PRODUCE (undated), https://specialtyproduce.com/produce/Red_Bananas_549.php (last updated Nov. 22, 2023).
[159] *See* Wiggins, *supra* note 148.
[160] *Gold Standard*, def. 2, MERRIAM WEBSTER, https://www.merriam-webster.com/dictionary/gold%20standard (last updated Mar. 30, 2024); *see Benchmark*, def. 1, MERRIAM WEBSTER (undated), https://www.merriam-webster.com/dictionary/benchmark (last updated Mar. 8, 2025); Curtis Northcutt et al., *Pervasive Label Errors in ML Datasets Destabilize Benchmarks, in* 35TH CONF. ON NEURAL INFO. PROCESSING SYS., TRACK ON DATASETS & BENCHMARKS 1, 1–2 (2021) (archived at ARXIV, https://arxiv.org/pdf/2103.14749.pdf) [hereinafter Northcutt et al., CONF. PAPER]. *See also* Curtis Northcutt et al., *Pervasive Label Errors in ML Datasets Destabilize Benchmarks*, L7 (Mar. 29, 2021), https://l7.curtisnorthcutt.com/label-errors (blog post summarizing & providing images & visualizations of research).
[161] Northcutt et al., CONF. PAPER, *supra* note 160, at 2.
[162] *Id.*
[163] *See id.*
[164] *Id.* at 2, 14.
[165] *Id.*

to improve the consistency and quality of annotations, including reducing the injection of annotation bias into AI systems.[166]

D. Learning Biases

Learning bias is the category of biases that arise as the machine learning system "finds" or "learns" its model.[167] Among the AI biases under this learning bias umbrella are at least seven: data dredging biases that result as AI system developers engage in probability or so-called *p*-hacking[168] by making outcome-driven learning modifications across a range of statistical metrics[169]; detection bias[170]; feature bias and, a subspecies, feature selection bias[171]; hyperparameter bias[172]; inductive bias[173]; and omitted variable bias.[174] Here, the Article discusses feature bias as its exemplar AI learning bias after briefly laying out the necessary background.

1. Of AI models, features, and functions

An AI model is a mathematical representation that approximately describes the correlations that are found to exist within the AI system's

---

[166] *See*, *e.g.*, Theresa Ann Wilson, Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States 24–25, 28–33, 180 (2008) (Ph.D. thesis, University of Pittsburgh), https://d-scholarship.pitt.edu/7563/1/TAWilsonDissertationApr08.pdf.
[167] *See* Hellström et al., *supra* note 24, at 2, 4–6.
[168] NIST REP'T, *supra* note 30, at 27, 50.
[169] *See* Adrian Erasmus et al., *Data-dredging Bias* (2020), *in* OXFORD BIAS CATALOGUE, *supra* note 27; NIST REP'T, *supra* note 30, at 27 & 50.
[170] *See* NIST REP'T, *supra* note 30, at 50. *See*, *e.g.*, '989 Patent, *supra* note 133, at 21.
[171] *See* Hellström *et al.*, *supra* note 24, at 3; Muhammad Bilal Zafar et al., *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment*, *in* PROC. OF 26TH INT'L CONF. ON WORLD WIDE WEB 1171, § 5, 1175–76 (Apr. 2017), (https://arxiv.org/pdf/1610.08452.pdf); Xiang, *supra* note 119, at 666–71. *See*, *e.g.*, Anne Washington, *How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate*, 17 COLO. TECH. L. J. 131, 152–53 (2019).
[172] *See* Hellström et al., *supra* note 24, at 2 (hyphen omitted); Jason Brownlee, *Understand the Impact of Learning Rate on Neural Network Performance,* MACHINE LEARNING MASTERY (Sept. 12, 2022), https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/.
[173] *See* Hellström et al., *supra* note 24, at 2; Loza de Siles, *supra* note 24, at 1402–04; Mitchell, *supra* note 24, at 1; Łukasz Gebel, *Why We Need Bias in Neural Networks*, TOWARDS DATA SCIENCE (Aug. 21, 2020), https://towardsdatascience.com/why-we-need-bias-in-neural-networks-db8f7e07cb98.
[174] *See In the Matter of Off. of Fed. Cont. Compliance Prog's, U.S. Dep't of Lab., v. Oracle America, Inc.,* 17-OFC-00006, 2020 WL 6112340 (Sept. 22, 2020) (citing hearing testimony); Takshi, *supra* note 120, at 248 n.162; '989 Patent, *supra* note 30, at 21.

intended subject domain, or the system's "world." Charmingly and aptly described, a model is the "toy version" of the real world.[175] An AI model "is an algebraic statement of how th[at toy] world works."[176] Because the model's world is a toy version, that modeled world differs from the real world. As distinguished statisticians and professors George E. P. Box and Norman R. Draper taught, some models are useful, *all* models, all of them, are wrong. [177] Models do not reflect the truth, and their approximate nature "must always be borne in mind."[178]

Within an AI model's toy world, patterns exist within the subject data sets correlate in some fashion to the AI system's target output. A target output might a juvenile person's predicted proclivity to commit violence at some point in the future. The correlation between features within the data and the target output is called a function. For example, the AI model may correlate whether a child has an attention-deficit or hyperactivity disorder[179] to a computed prediction of his or her likelihood to engage in some violent behavior in future.[180]

The patterns do not emerge from the dataset as a whole. They emerge from particular data, and the numerical and categorical values for those data, within that dataset. These particular data that are relevant by virtue of some correlation with system's objective are called "features," or, synonymously, "attributes." [181] A feature is an independent input variable within the dataset that correlates in some way, to some degree, and in some pattern of combination with other features, collectively, the "feature" set, to the desired AI output, or the "dependent" output variable.[182] In the above example, the applicable data element(s) within the data set might be labeled along the lines of "neurological diagnoses,"

---

[175] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 26 (2016).
[176] Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOCIO. METHODS & RES. 1, 3 n.2 (2021), https://doi.org/10.1177/0049124118782533.
[177] *See* GEORGE E. P. BOX & NORMAN R. DRAPER, EMPIRICAL MODEL-BUILDING AND RESPONSE SURFACES 424 (1987).
[178] *Id.*
[179] *See* Paul Simpson, *Studies on Violence Risk in Youth* 1, 7 (2014) (summarizing studies underlying Structured Assessment of Violence Risk in Youth, or SAVRY, system) (on file with author).
[180] *See* Loza de Siles, *supra* note 24, at 1424–25 (2021).
[181] *Machine Learning Glossary*, GOOGLE DEVELOPERS ("feature" entry, https://developers.google.com/machine-learning/glossary#feature); "attribute" entry, https://developers.google.com/machine-learning/glossary#attribute) (last updated Mar. 11, 2024).
[182] Jason Brownlee, *Machine Learning Terminology from Statistics and Computer Science*, MACHINE LEARNING MASTERY (Mar. 19, 2016), https://machinelearningmastery.com/data-terminology-in-machine-learning/.

and where the values for those elements are "Yes" and "ADD" or "ADHD," those features within the data are correlated within the AI model's toy world as correlated to an increased risk of the child's future violence.

The iterative processes by which AI models are "learned" or "found," in the vernacular, and later selected and optimized involves the identification of correlating features; the delineation of the feature set; and the applications of weights to features so as to express their relative importances [183] and, thereby, hopefully improve their contribution toward the AI model's function.[184] In the example, the AI system's output is a conditional probability [185] where that output is presumed to be conditionally "true" vis-à-vis the target function of AI model's toy world.

2. Feature bias

Feature bias is the group of biases involving the features that are used within the subject dataset(s) to learn AI models.[186] Feature bias may manifest in multiple ways. One type of feature bias arises when two conditions accrue. The first condition is that certain features within the feature set are only moderately correlated to the output variable.[187] Relative to other features within the feature set, these features are "weak" in terms of their predictive power toward the desired output.[188] The second condition is that insufficient data are used to train the subject system.[189] As the combined result of these two conditions, the weight attributed to these weakly predictive features is or may over time become

---

[183] *See Machine Learning Glossary*, GOOGLE DEVELOPERS, https://developers.google.com/machine-learning/glossary#variable-importances ("variable importances" entry) (last updated Nov. 14, 2023).
[184] Brownlee, *supra* note 182 ("[I]n the phrasing of the prediction problem[,] the output is dependent or a function of the input or independent variables.")
[185] *See Machine Learning Glossary*, GOOGLE DEVELOPERS, (last updated Nov. 14, 2023), ("discriminative model" entry, https://developers.google.com/machine-learning/glossary#discriminative-model; & fig. accompanying "neural network" entry, https://developers.google.com/machine-learning/glossary#neural-network).
[186] Feature bias as under discussion here differs from that term used in other contexts elsewhere. *See* note 171 & accompanying text (feature & feature selection biases as types of learning biases). The term "feature bias" also has been discussed as to perceptions of defendants' facial features as "Afrocentric" and as presenting as an aspect of stereotype bias in criminal proceedings. *See generally*, *e.g.*, Amanda M. Petersen, *Complicating Race: Afrocentric Facial Feature Bias and Prison Sentencing in Oregon*, 7 RACE & JUSTICE 59 (2017), https://doi.org/10.1177/2153368716663607.
[187] *See* Klas Leino et al., *Feature-Wise Bias Amplification*, ARXIV 1 (last updated Oct. 21, 2019) (presented at Seventh Int'l Conf. on Learning Representations (2019)), https://arxiv.org/abs/1812.08999.
[188] *Id.*
[189] *Id.*

outsized in comparison to their predictive value in actuality.[190]

Another species of feature bias occurs when certain features that are associated with the "ingroup" are more heavily weighted and, therefore, resulting in the biased favoring of that privileged ingroup as compared to the unprivileged "outgroup."[191] Amazon's AI recruitment tool did just that.[192] Amazon sourced the dataset for the tool from 50,000 resumes received by the company, overwhelmingly from men,[193] for software engineering and other technical positions during a ten-year period.[194]

Leaving aside the matter of data biases, feature bias entered into play in this Amazon example. Men constituted the privileged group and women the unprivileged. Among the categorical features that became privileged as Amazon's system learned its model were textual representations of stereotypical masculinity, such as the resumes' uses of

---

[190] *See id.* at 1, 5; *see also id.* at 4-5 (discussing "feature asymmetry").

[191] *Ingroup*, *APA DICTIONARY OF PSYCHOLOGY*, (last updated Apr. 19, 2018), https://dictionary.apa.org/ingroup (defining "ingroup" generally as "any group to which one belongs or with which one identifies, but particularly a group judged to be different from other groups (outgroups)."); *id.* at *Outgroup* (last updated Apr. 19, 2018), https://dictionary.apa.org/outgroup (defining "outgroup" generally as "any group to which one does not belong or with which one does not identify."). *See generally* Saad Ahmed et al., *Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach*, IEEE PROC. OF 2021 6TH INT'L CONF. ON INVENTIVE COMPUTATION TECH'S 557 (2021), https://doi.org/10.1109/ICICT50816.2021.9358507.

[192] *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, THOMSON REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. In 2014, Amazon began developing the tool and halted its use in early 2017, an estimated two years after Amazon discovered the system's significant gender bias. *See id.*

This discussion concentrates on the Amazon system's gender bias as an illustration of feature bias. There were other prediction flaws with the system, however, *see id.*, or with these types of systems generally, *see* Rachel Goodman, *Why Amazon's Automated Hiring Tool Discriminated Against Women,* AMER. CIV. LIBERTIES UNION (Oct. 12, 2018), https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against; Matt Gonzales, *AI-Based Bias a Hot Topic of Discussion During EEOC-Led Meeting*, SOC'Y FOR HUMAN RESOURCE MGMT. (Oct. 7, 2022), https://www.shrm.org/resourcesandtools/hr-topics/behavioral-competencies/global-and-cultural-effectiveness/pages/ai-based-bias-a-hot-topic-of-discussion-during-eeoc-led-meeting.aspx.

[193] *See* Goodman, *supra* note 192.

[194] Dastin, *supra* note 192; *see* Goodman, *supra* note 192 (software engineering & other technical positions). Men greatly predominate within these types of jobs and within the technology industry generally. *See Female Representation in Technology Organizations in 2021, by Selected Countries,* STATISTICA (Aug. 2021), https://www.statista.com/statistics/1256204/representation-of-gender-tech-by-country/; Dastin, *supra* note 192, at fig's (attributing "Global Headcount" & "Employees in Technical Roles" figures to Han Huang, Reuters Graphics (dating source(s) as since 2017)).

the words "executed" and "captured."[195] Conversely, the system's model learned to disfavor candidates whose resumes included the word "woman" and its variants or indicated that they had attended one of at least two all-women colleges and universities.[196]

This favoring and disfavoring between the ingroup and the outgroup, respectively, occurs, at least in part, by the assignment of greater, that is, more positively correlated, relative weights to features associated with the former than with the latter.[197] In Amazon's case, the feature bias toward male candidates and against their female competitors for these technical positions became so "hardened" within the AI model so as to be irremediable.[198]

E. Model Biases

Model biases are the category of AI biases observed when evaluating the outputs of finalized AI models, including, for example, classification differences between groups of AI subjects upon which those models compute.[199] There are at least six AI model biases, namely:

---

[195] *Id.*

[196] *Id.*; *accord id.* (women's colleges); *see Women's Colleges in the United States*, WIKIPEDIA (last updated Feb. 25, 2024), https://en.wikipedia.org/wiki/Women%27s_colleges_in_the_United_States.

[197] The privileged group may be comparatively monolithic and more homogeneous as compared to the unprivileged group, a characteristic that likewise may cut in favor of the privileged group. *See* Khari Johnson, *Pymetrics Open-sources Audit AI, an Algorithm Bias Detection Tool*, VENTUREBEAT (May 31, 2018), https://venturebeat.com/ai/pymetrics-open-sources-audit-ai-an-algorithm-bias-detection-tool/ ("[T]op performers at some companies can be overly represented by a single, homogeneous demographic group."); *accord generally* Poornima Nataraja & Bharathi Ramesh, *Machine Learning Algorithms for Heterogenous Data*, 10 INT'L J. COMP. ENG'G & TECH. 1-3 (2019), https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_10_ISSUE_3/IJCET_10_03_002.pdf.

[198] *See* Goodman, *supra* note 192.

[199] Hellström et al., *supra* note 24, at 5, 7. These authors liken "model bias" to, but conceptualize it more broadly than, "algorithmic bias," a term frequently discussed and usually connoting social justice concerns. *See* Hellström et al., *supra* note 24, at 4. They also include statistical measurements of bias within the model bias category. *See generally also* Mayson, *supra* note 44. Because these metrics, however, are not types of biases, but rather bias metrics, this Articles excludes them from its model biases category.

feedback loop bias[200]; linking bias[201]; model selection bias[202]; temporal bias[203]; misclassification bias[204]; and uncertainty bias.

To discuss one, **uncertainty bias** is a type of model bias indicative of the degree of accuracy, or lack thereof, in the AI system's results.[205] Uncertainty bias may arise directly, but also may be produced as the result of other AI biases, *e.g.*, sampling bias.[206] Uncertainty bias is closely associated with a statistical concept known as a "confidence value." Confidence value is statistical metric that reflects the degree to which a computed result as likely to be objectively accurate.[207] The higher confidence value, the lesser degree of uncertainty there is calculated to be in the accuracy of the result. Conversely, the lower the confidence value, the greater the degree of uncertainty. In general, then, the aim is for the confidence value to be high so that the uncertainty as to the accuracy of the result will be low. As a practical matter, the ideal of 100% confidence in a predictive result is not possible or reasonably so, however.[208]

For instance, where the subject dataset under-represents non-majority populations, uncertainty bias undermines the confidence that should be placed in the AI system's predicted classifications.[209] Such

---

[200] *See* NIST REP'T, *supra* note 30, at 51; SCIENTIFIC EVIDENCE, *supra* note 90, at 872; Sharona Hoffman & Andy Podgurski, *Artificial Intelligence and Discrimination in Health Care*, 19 YALE J. HEALTH POL'Y & ETHICS 1, 7, 15–16 (2020).
[201] *See* Mehrabi et al., *supra* note 30, at 115-16; NIST REP'T, *supra*, at 52.
[202] *See* Alexis Bogroff & Dominique Guegan, *Artificial Intelligence, Data, Ethics[:] An Holistic Approach for Risks and Regulation* 18-19 (July 9, 2019), Univ. Ca' Foscari of Venice, Dept. of Econ. Rsch. Paper Series No. 19, 18-19, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3419360; NIST REP'T, *supra* note 30, at 53.
[203] *See* Mehrabi et al., *supra* note 30, at 8; NIST REP'T, *supra* note 30, at 54.
[204] *See generally* Kevin Kloos et al., *Comparing Correction Methods to Reduce Misclassification Bias*, *in* ARTIFICIAL INTELLIGENCE & MACHINE LEARNING: 32ND BENELUX CONF., BNAIC/BENELEARN 2020 REVISED SELECTED PAPERS, 64 (Leiden, Netherlands, 2020) (M. Baratchi et al., eds., 2021), https://doi.org/10.1007/978-3-030-76640-5_5; *see*, *e.g.*, '989 Patent, *supra* note 133, at 22, col. 20 & 23, col. 21.
[205] *See* Mounib Khanafer & Shervin Shirmohammadi, *Applied AI in Instrumentation and Measurement: The Deep Learning Revolution*, 10 IEEE INSTRUMENTATION & MEASUREMENT MAG. 13 (Sept. 2020), https://ieee-ims.org/sites/ieeeims/files/2021-01/Deep%20Learning%20Topical%20Guide.pdf; Hellström et al., *supra* note 24, at 2. NIST describes uncertainty bias differently than do Hellström and his co-authors. *Compare* NIST REP'T, *supra* note 30, at 54 *with* Hellström et al., *supra* note 24, at 2.
[206] *See* Hellström et al., *supra* note 24, at 2.
[207] For instance, a weather prediction may call for an 80% chance of thunderstorms. A confidence value reflects how statistically sure one may be that this 80% prediction is accurate. If the confidence value is 100%, then it is statistically certain that, indeed, there is an 80% chance of thunderstorms. Conversely, if the confidence value is only 50%, then the 80% prediction equally may or may not be accurate.
[208] The ideal confidence value of 100% is often not reasonably achievable.
[209] *See* Hellström et al., *supra* note 24, at 2.

uncertainty bias also taints the confidence that should be instilled in the thresholds that are chosen to demarcate the system's classes.[210] For example, children who are subjected to use of the Structured Assessment of Violence Risk in Youth, or SAVRY, system are deemed to have elevated risks of future violence if their characteristics, and characteristics of those around them, are collectively classified through SAVRY as "medium" or "high."[211]

Held to account under the reliability requirement under the Federal Rules of Evidence and *Daubert*, however, the SAVRY system as applied in *In re* T.K. has failed.[212] In that case, Judge Robert Okun of the District of Columbia Superior Court excluded the SAVRY system's results from reference or evidentiary use.[213] The Court did not use the term "uncertainty bias" or elaborate the basis for its ruling much beyond its agreement that "Respondent has raised a number of valid criticisms" with the SAVRY system as applied.[214] Among those criticisms were the failure to test the SAVRY system and the absence of any error rate information.[215] At the bottom and at least in part, the Court's order stands for the propositions that some degree of uncertainty bias existed in the application of the SAVRY model in T.K.'s case and that, absent the testing, quantification, and acceptability of such bias, the use of the system was legally impermissible.

AI models produce interim or final outputs that are based on probabilities, and all probabilities are, by nature, not deterministic. Because AI models are probabilistic and for other reasons, all AI models

---

[210] *See id.* at 2 & n.5.

[211] *See* Motion to Exclude Results of the Violence Risk Assessment and all Related Testimony and/or Allocution Under FRE 702 and Daubert v. Merrell Dow Pharmaceuticals, In *re* T.K. 7, 10 (D.C. Super. Ct. Feb. 5, 2018) [hereinafter SAVRY Motion], *linked in* AI NOW INST., CENTER ON RACE, INEQUALITY, AND THE LAW & ELECTRONIC FRONTIER FOUNDATION, LITIGATING ALGORITHMS: CHALLENGING GOVERNMENT USE OF ALGORITHMIC DECISION SYSTEMS 13 (Sept. 2018), https://ainowinstitute.org/litigatingalgorithms [hereinafter 2018 LITIGATING ALGORITHMS].

[212] *See* Order, In re T.K. (D.C. S. Ct., Mar. 15, 2018) [hereinafter SAVRY Order] (granting, in part, Respondent's Motion to Exclude Results of the Violence Risk Assessment and all Related Testimony and/or Allocution Under FRE 702 and *Daubert v. Merrell Dow Pharmaceuticals*), *linked in* 2018 Litigating Algorithms, *supra* note 211; Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993).

[213] For procedural and corresponding evidentiary reasons, the Court did not opine on the inherent validity of the SAVRY system. *See* SAVRY Order, *supra*, at 7-8. Correspondingly, the Court's order did not reach to other youths exposed to the SAVRY system during the District of Columbia's then-14 year use of the system and what may be its subsequently continuing use.

[214] *Id.* at 8.

[215] *See generally* SAVRY Motion, *supra* note 211; *id.* at 7-9 (biases & errors, particularly as to youth of low socioeconomic classes).

are wrong,[216] that is, their functions deviate from the objective truth. Because all AI models are wrong, all exhibit the errors that constitute uncertainty bias.

The law, however, has had a long and checkered history of accepting, almost without question, unknown but unacceptably high degrees of uncertainty bias in cases where innocent defendants are convicted and sentenced to death.[217] As abominable as that is,[218] the scale and pervasiveness with which AI systems and their uses may distribute the impacts of uncertainty bias errors in AI models also may be terrible in the aggregate. Conversely, knowledge and study of uncertainty biases in machine intelligence models may promote a thorough re-examination of the law's acceptance of these biases more broadly and particularly as present in human decision-making.

F. Use Biases

Use biases are the errors that result from the AI system being deployed, implemented, or used in a manner or for an application other than that for which it was designed and developed, that is, other than

---

[216] *See* note 177 & accompanying text.

[217] *See*, *e.g.*, Rita James Simon, *"Beyond a Reasonable Doubt"—An Experimental Attempt at Quantification*, 6 THE J. OF APPLIED BEHAV. SCI. 203, 203 (1970), doi:10.1177/002188637000600205; Jon O. Newman, *Quantifying the Standard of Proof Beyond a Reasonable Doubt: A Comment on Three Comments*, 5 L., PROBABILITY & RISK 267, 267-68 (2006), https://academic.oup.com/lpr/article-pdf/5/3-4/267/2711717/mgm010.pdf; Katie Evans et al., *Distributions of Interest for Quantifying Reasonable Doubt and Their Applications*, NAT'L SCI. FOUND. 14 (2006)., https://www.valpo.edu/mathematics-statistics/files/2015/07/Distributions-of-Interest-for-Quantifying-Reasonable-Doubt-and-Their-Applications.pdf.

[218] *See*, *e.g.*, Paula Christian, *Man Set Free After 28 Years in Prison Dies Before Wrongful Conviction Suit Against Newport Ends*, WCPO ABC 9 CINCINNATI (Mar. 2022), https://www.wcpo.com/news/local-news/i-team/man-set-free-after-28-years-in-prison-dies-before-wrongful-conviction-suit-against-newport-ends. Some 187 persons have been exonerated from death row since 1976 from more than half of the United States, each of them having spent more than 11 years on death row for crimes they did not commit. WITNESS TO INNOCENCE, *February 2021 Special Report: The Innocence Epidemic* (Feb. 2021) (citation to Death Penalty Info. Ctr. report omitted), https://www.witnesstoinnocence.org/innocence. For every 8 persons judicially killed by the state, 1 has been exonerated. *Id.*

A 2014 study published in the Proceedings of the National Academy of Sciences determined that at least 4.1% of people on death row in the United States are falsely convicted and thus innocent. Samuel R. Gross et al., *Rate of False Conviction of Criminal Defendants who Are Sentenced to Death*, Proc. 111 Nat'l Acad. of Sci. 7230, 7230 (2014)., https://doi.org/10.1073/pnas.1306417111. As of 2021, there were 2436 women and men on death row or subject to resentencing to death in the United States. DEATH PENALTY INFO. CTR., *Size of Death Row by Year* (undated), https://deathpenaltyinfo.org/death-row/overview/size-of-death-row-by-year (last visited Mar. 10, 2025). Using the conservative estimate published by the National Academy of Sciences, at least 100 of these people are innocent. *Id.*

directed toward its target concept.[219] Here, the term "use" is broadly conceived. This "use" spans the AI lifecycle and all the processes and activities from the conclusion of procurement and onward through implementation, integration, training, placement into production for operational use, those operations, and further through the sunsetting of the AI system, the user organization's potential migration to another system, and on through the long tail of archival practices, litigation holds, ongoing discovery in litigation, government records and transparency laws, and accounting and other recordkeeping requirements.[220] There are at least seven AI use biases, as follow: deployment bias[221]; emergent bias[222]; mode confusion bias[223]; off-label use bias; omission bias[224]; reliance bias[225]; and user interaction bias.[226]

Off-label use bias describes the errors that result when AI system users employ those systems for other than the target concept for which those systems are designed, developed, and marketed.[227] Such misuse has been analogized to off-label use of pharmaceuticals, that is, uses for which the drug has not been approved as safe and effective for

---

[219] *See* Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 153 (2017), https://doi.org/10.1089/big.2016.0047. To avoid confusion, note that the NIST special report defines the narrower term "deployment bias" as bias created by use generally. *See* NIST REP'T, *supra* note 30, at 50.

[220] *See generally*, *e.g.*, Aran Davies, *What is the AI Software Development Life Cycle?*, DEVTEAM.SPACE (undated), https://www.devteam.space/blog/ai-software-development-life-cycle-explained/ (last visited Mar. 10, 2025); Christopher S. Penn, *How to Get Started with Machine Learning and AI*, CHRISTOPHER S. PENN (Feb. 2, 2017) (lifecycle fig.), https://www.christopherspenn.com/2017/02/how-to-get-started-with-machine-learning-and-ai/.

[221] *See* Chouldechova, *supra* note 219, at 153. To avoid confusion, note that NIST defines the narrower term "deployment bias" as bias created by use generally. *See* NIST REP'T, *supra* note 30, at 50.

[222] See Batya Friedman & Helen Nissenbaum, Bias in Computer Systems, 14 ACM TRANS. INF. SYST. 330, 335 (July 1996) https://doi.org/10.1145/230538.230561; NIST REP'T, *supra* note 30, at 50.

[223] *See* NIST REP'T, *supra* note 30, at 52.

[224] *See*, *e.g.*, '989 Patent, *supra* note 133, at 19, col. 13.

[225] *See generally, e.g.*, Cl'elie Amiot et al., Whom Do We Trust?: A Comparative Study on Reliance between Chatbot and Human Assistance (Oct. 5, 2023) (Univ. of Lorraine working paper), https://hal.univ-lorraine.fr/hal-04229730/file/Whom_Do_We_Trust.pdf; Jennifer Ross, Moderators Of Trust And Reliance Across Multiple Decision Aids, iii-iv (2008) (Ph.D. dissertation, Univ. of Cent. Fla.), https://core.ac.uk/download/pdf/236258794.pdf.

[226] *See* Baeza-Yates, *supra* note 132, at 58-60.

[227] *See generally*, *e.g.*, Erin Collin, *Punishing Risk*, 107 GEO. L.J. 57 (2018).

use by the regulatory authority.[228] Drug manufacturers generally do not face liability for harms that result from physicians' prescription of drugs for off-label uses because those physicians are deemed learned intermediaries.[229] Because they are educated and experienced so as to know their business, this learned intermediary doctrine holds physicians, and not drug manufacturers, accountable for harms that result from off-label uses of those pharmaceutical products.

Off-label uses of AI systems occur in the absence of governance or legal requirements otherwise. For example, courts have used risk predictor AI systems to make sentencing or imprisonment condition decisions where, in at least one challenged case, the system's target concept was directed toward and its use expressly limited to identifying offenders for intervention services and risks that may impact upon their supervision as to those services.[230]

Unlike as in the off-label uses of pharmaceuticals, however, the likely overwhelming majority of individuals who engage in and expose their human computational subjects to off-label uses of AI systems are a far cry from AI-learned intermediaries. Further, the errors associated with automation and other cognitive biases compound the errors attributable to off-label use bias.[231] Nevertheless, these unlearned AI intermediaries have not been held to account in any exemplary or appreciable sense, and neither have the purveyors of such misused AI systems, despite the absence of a shielding learned intermediary defense.[232]

IV. AI BIAS MECHANISMS

---

[228] The term "off-label use" originated to describe the use of pharmaceuticals outside the scope of their regulatorily-approved and thus-labelled use. Before pharmaceutical drugs may be legally sold, they must be determined to be safe and efficient in the context of the manufacturer's proposed use, and meet other regulatory requirements. *See generally* U.S. FOOD & DRUG ADMIN., *Development & Approval Process | Drugs* (last updated Aug. 8, 2022), https://www.fda.gov/drugs/development-approval-process-drugs.

[229] *See generally* Rebecca Dresser & Joel Frader, *Off-label Prescribing: A Call for Heightened Professional and Government Oversight*, 37 J. L. MED. & ETHICS 476 (2009), doi: 10.1111/j.1748-720X.2009.00408.x.

[230] *See*, *e.g.,* State v. Loomis, 371 Wis. 2d 235, 246 (2016) (quoting Presentencing Investigative Report' express notice of intended use of Northpointe's COMPAS system & caution against unintended uses thereof).

[231] *See* Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207, 207 (1996) ("Beliefs [that] lawyers hold about computers, and predictions they make about new technology, are highly likely to be false. . . . The blind are not good trailblazers.").

[232] *Compare* Emile Loza De Siles, *FDA Regulation of Internet Pharmaceutical Communications: Strategies for Improvement*, 55 FOOD & DRUG L.J. 269, n. 30, 48, 90-91, 103 & accompanying text (2000).

This Section describes the ways in which AI biases originate and interact within the lifecycle of AI as human-machine enterprise and maps those mechanisms to that lifecycle. This mapping of AI bias mechanisms helps to identify loci of control, governance, and potential regulation as to AI biases, collectively "AI governance control points."

As a preliminary matter, note that multiple AI biases may and arguably must simultaneously exist within the AI lifecycle. These multiple biases also may interdigitate with one giving rise to or exacerbating another and so on through the lifecycle, which, in the end, contains a cascade of AI biases.

A. Bias Injection

This Article is the first law scholarly work to closely consider and discuss how AI bias injection works.[233] A look toward cybersecurity provides a well-established and similar phenomenon, however. There, bias, or error, injection is a well-understood weaponization mechanism by which bad actors execute cyberattacks. There, attackers seek to increase system errors, exfiltrate data, or disable or destroy the attacked system altogether by injecting false data or other biasing information or malicious code into the critical infrastructure or other targeted systems.[234] A critical feature of this attack strategy is that the error and its injection are planned to remain obfuscated and undetectable for as long as possible.[235] Under this covert cover, nefariously injected data, information, or code persist, remaining undiscovered and doing their dirty work sometimes for years.[236] A similar attack strategy is now being executed against large language AI models ("LLMs") using prompt injection attacks so as to manipulate or otherwise corrupt the targeted LLM's output.[237]

Bias injection is the mechanism by which biases arise within or enter into processes within the AI lifecycle. By identifying and understanding the various types of AI biases and situating them within

---

[233] One article provides a short discussion of data bias injection. *See* Yafit Lev-Aretz, *Data Philanthropy*, 70 HASTINGS L.J. 1491, 1519 (2019).
[234] *See* Jezdimir Milošević et al., *Analysis and Mitigation of Bias Injection Attacks Against a Kalman Filter*, INT'L FED. OF AUTOMATIC CONTROL PAPERSONLINE 50-1, 8393, 8393 (2017), https://doi.org/10.1016/j.ifacol.2017.08.1564.
[235] *See id.*
[236] *See id.*
[237] *See* Rich Harang, *Securing LLM Systems Against Prompt Injection*, NVIDIA TECH. BLOG (Aug. 3, 2023), https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/.

the AI lifecycle, it is possible to identify the points at which different biases do or may enter into one or more lifecycle processes. Call these "AI bias injection points."[238]

1. Single-Point Bias Injection

Some types of AI biases may arise or be introduced at only one phase of the AI lifecycle or during one process within one phase. These AI biases could be considered to occur through a single injection point mechanism, and consequently, their governance and control would focus on that "single" point. Hyperparameter bias is one such AI bias that arises or occurs through single-point injection. Hyperparameters are adjustable aspects of the machine learning process, and the learning rate is one of these hyperparameters. Remember that this learning process finds correlations between features within the training data and the target concept of the AI system. Those correlations result in various models, and after numerous iterations, weighting adjustments, and so on, one is ultimately selected and optimized for deployment.

Learning rates may be increased or decreased, and there are trade-offs to making those adjustments. Generally speaking, with a slower learning rate, the accuracy of the correlations improves, but the time and costs of learning the model increase. Conversely, with a faster learning rate, the learning costs decrease, but so too does the accuracy. Imagine being a passenger in a car and driving alongside a fenced field of corn or other crop rows running perpendicular to the road. As the speed of the car increases, the fence posts and crop rows begin to blur in the passenger's vision. As the speed decreases, the posts and the crop rows become more discretely discernable within that vision. In this example, the blurring or the degree to which the posts and crop rows cannot be individually seen in their true forms constitutes a bias, or error, in the vision where the goal, or the target concept, is to see each distinctively. Similarly, adjustments to the learning rate and other hyperparameters may produce biases. Because hyperparameters are associated only with the learning phase of the AI lifecycle, hyperparameter bias has a single

[238] Emile Loza de Siles, *Disaggregating Artificial Intelligence Biases: A Law and Systems Engineering Approach for AI Governance and Regulation*, *in* RESEARCH HANDBOOK ON THE LAW OF ARTIFICIAL INTELLIGENCE [29]-[33] (2d.,Woodrow Barfield & Ugo Pagallo, eds.) (Edward Elgar Pub'g, forthcoming 2024) (on file with author).

point of injection into that lifecycle.[239]

2. Multipoint Bias Injection

Some types of bias, however, may be injected at more than one phase of the AI lifecycle. Collectively, AI biases that have the potential to function by this mechanism are coined here as multipoint injectable biases. Within the class of AI biases, there are some further variations in the mechanism by which certain biases may multiply arise or be injected into the AI lifecycle.

a. Range-Restricted Multipoint Bias Injection

To discuss two,[240] one subclass of these biases may be injected only in two or more phases or a certain span of phases, or a restricted, but not necessarily sequential range, within the AI lifecycle. Building upon this Article's earlier discussion of the SAVRY violence risk predictor,[241] various data biases are reflected in that AI human-machine

---

[239] For clarity, this Article presents the AI enterprise lifecycle as a flat sequence without hierarchy and then maps biases to phases within that sequence. Note that, in truth, there may be what amount to multiple layerings of AI lifecycles that result in some form of hierarchy. For example, data mining and more broadly data curation processes fall within the "input data" phase of this Article's AI enterprise lifecycle. These data-related processes may incorporate and themselves depend upon machine learning subprocesses. *See generally*, *e.g.*, Michael Stonebraker et al., *Data Curation at Scale: The Data Tamer*, Remarks at the 6th Biennial Conf. on Innovative Data Sys. Rsch., in Asilomar, Cal. (Jan. 6–9, 2013), https://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper28.pdf; Jan. N. van Rijn & Frank Hutter, *Hyperparameter Importance Across Datasets*, *in* PROC. OF THE 24TH ASS'N COMP. MACHINERY ("ACM") SPEC. INTEREST GROUP ON KNOWLEDGE DISCOVERY & DATA MINING ("SIGKDD") INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 2367 (July 2018), https://doi.org/10.1145/3219819.3220058; *see also* ACM SIGKDD, *SigKDD* (undated), https://kdd.org/about (last visited Mar. 10, 2025). The injection of hyperparameter bias within such machine learning subprocesses may go on to affect the "parent" input data phase of and on through the top-level AI enterprise lifecycle. This may occur even where machine learning subprocesses, with complex and finely tuned hyperparameters, are applied toward the mitigation of biases in datasets. *See*, *e.g.*, Pranita Patil & Kevin Purcell, *Decorrelation-Based Deep Learning for Bias Mitigation*, 14 FUTURE INTERNET 1, 1-2 (Daniel Gutiérrez Reina, ed., 2022), https://doi.org/10.3390/fi14040110.

[240] In addition to the two subclasses of multipoint injectable biases mentioned in the text, there is another subclass, coined here "chimera biases," that vary in their manifestations and implications, technical, legal, or both, depending upon at which multiple points they arise or are injected into the AI lifecycle. The exploration of chimera biases is reserved for a future work.

[241] *See supra* note 210-15 & accompanying text.

enterprise[242] and illustrate these range-restricted multipoint injectable biases.

Briefly, the SAVRY violence predictor model is based upon various studies that purport to correlate certain immutable characteristics of the subject juvenile, a child or youth, and that juvenile's family, school, and community circumstances. [243] Medical conditions are among these immutable characteristics: attention-deficit disorder, hyperactivity disorder, and, presumably, attention deficit-hyperactivity disorder.[244] These neurological conditions constitute developmental disabilities and, further, constitute disabilities that qualify the people who have these conditions for protection under the Americans with Disabilities Act people having these conditions.[245]

One study underlying the SAVRY model reportedly found that youth with these medical conditions were more than four times more likely than youth in the study's control group to be arrested (46%, as compared to 11%); almost four times for likely to be arrested on suspicion of violent crimes (36%, as compared to 9%); and twenty-two times more likely to be incarcerated (22%, as compared to 1%).[246]

Given the legal status of ADD, HDD, and ADHD as disabilities for which those afflicted with them are protected, the inclusion of any such data as training data, which are implicated in the data, learning, and modeling phases of the AI lifecycle, may be illegal, unfair, given their immutability, or, at best, erroneous. Because bias is an error, then this inclusion and use of these data constitutes one or more types of AI data bias. This is not the end of this AI bias story with SAVRY, however.

The SAVRY predictor model, as developed or implemented, groups these medical conditions together as "Item #22." This Item #22 is used, that is, presumably scored, in three risk predictive categories: (1) "disruptive/behavioral problems"; (2) "mental health / emotional stability"; and (3) "education/employment." [247] People having ADD,

---

[242] *See generally* Simpson, *supra* note 179; John S. Ryals, Jr., Jefferson Parrish Dep't of Juvenile Svcs., Screening & Assessment Manual: Structured Decision-Making Across Jefferson Parish Juvenile Justice System 1, 11-13 (3rd ed., 2019) [hereinafter SAVRY Manual], https://jefferson-parish-government.azureedge.net/Screening%20and%20Assessment%20Manual%202019%20(FINAL).pdf.
[243] *See generally* Simpson, *supra* note 179.
[244] *See id.* at 7.
[245] *See generally* Office for Civil Rights, U.S. Dep't of Educ., *Know Your Rights: Students with ADHD* (July 2016), https://www2.ed.gov/about/offices/list/ocr/docs/dcl-know-rights-201607-504.pdf; Clifford M. Koen. Jr. et al, *Attention Deficit Disorder and the Americans With Disabilities Act: Is Anyone Paying Attention?*, 36 HEALTH CARE MGMT. 116 (Apr./June 2017), doi: 10.1097/HCM.0000000000000161.
[246] *See* Simpson, *supra* note 179, at 7.
[247] *See* SAVRY Manual, *supra* note 243, at 68 (App'x 6).

HDD, and ADHD conditions may experience behavioral difficulties; be labeled as disruptive, often due to bullying; and suffer education and employment impacts. Their challenges also can result in mental health challenges, and these, in turn, may produce some emotional lability. It seems patently unfair, if not outright illegal, however, to count, perhaps multiplicatively, the effects of and external reactions to their medical conditions against these individuals, even if there were or is some relevant and valid correlation between those conditions and violence risk. Consider, further, that many continue to be uneducated as to these conditions, and so label those afflicted with them as miscreants or other "problems."

Further still, ADD, HDD, and ADHD may be excluded from consideration during SAVRY scoring as largely immutable characteristics requiring some longitudinal perspective so as to determine the current effects of these conditions upon the subject juvenile.[248] This failure to consider these lifelong medical conditions are in fact immutable suggests that some additional biases are leveled against the affected individual, who is considered, abhorrently, to be grossly incalcitrant or somehow lacking in will or moral character for failing to affirmatively rewire his neurology. Cognitive biases, as well as social and cohort biases, clearly enter into the picture of AI biases here.

This, however, is still not the end of this AI bias story with SAVRY. Having used SAVRY for at least thirteen years,[249] Louisiana's Jefferson County Parish Department of Juvenile Services (in this subsection, "Department"), for example, provides the guidance to the people, currently probation officers, who collect data about the subject juvenile and his or her circumstances.[250] That guidance consists of: (1) a one-third page "script" that is stated as including questions, but does not[251]; (2) a one-page list of topics as to SAVRY risk predictive and protective factors[252]; (3) a Report to Court Outline ("Outline") form, which is not identified as related to SAVRY but contains questions intended to illicit data for use with SAVRY and likely one or more of the Department's other assessment tools[253]; and (4) a Service Referral Matrix, with which the

---

[248] *See id*. at 69 (App'x 7).
[249] *See* Kristina Childs et al., *Jefferson Parish Youth Outcomes Study* 2, LOUISIANA MODELS FOR CHANGE (2011) (indicating SAVRY had been implemented "since" 2007, and finding that only thirteen percent of youth treatments were "evidence-based"), https://sph.lsuhsc.edu/wp-content/uploads/2016/07/4BYouthOutcomes.pdf.
[250] *See* SAVRY Manual, *supra* note 243, at 11, 52, 69.
[251] The script consists of some minimal language meant to put the juvenile and his or her parents at ease. *See id.* at 52.
[252] *See id.* at 11-12.
[253] *See id*. at 55-67 (App'x 5).

Outline results are used to map SAVRY factors and risk levels as determined by the data collectors to the Department or other services.[254]

Muddled guidance aside, the Department releases data collectors, as they become "more comfortable" with SAVRY, to exercise their own discretion and proceed as they see fit, relegating the Outline to mere guidelines.[255] This deferral to the data collectors' discretion creates or increases the risk and perhaps even the certainty that multiple data biases will be injected in the use phase of the AI lifecycle, notwithstanding the question of a second injection of AI bias in the form of potentially illegally and almost certainly unfairly-included data about the juvenile's disabilities.

b. Globally Injectable Biases

A second subclass of multipoint injectable biases may arise at any or, for some biases, *every* point in the AI lifecycle, those being classified as **globally injectable biases**. Implicit bias is a cognitive bias that is present in most and likely all human minds, and one that has multiple characteristics as its foci.[256] Given its far-reaching and ubiquitous nature, implicit bias may arise globally across the AI human-machine lifecycle. For instance, implicit biases favoring younger adults over their older counterparts, Caucasian males over females, and people of color manifest themselves in the collection, composition, and choice of datasets used to train facial recognition systems.[257] Additional implicit biases may arise in the learning, modeling, and use phases. In this example, multiple implicit biases are injected into the AI lifecycle, each of them at a different point therein.

Given their complexity and implications, multipoint injectable biases and their corresponding mechanisms are particularly important to understand and for which to appropriately control and establish good AI governance. Think of exposure to multipoint injectable biases like exposure to ionizing radiation. Multiple exposures to even low-level radiation drive up the risks of cancer or other health impacts.[258] Stated

---

[254] *See id.* at 68 (App'x 6).
[255] *Id.* at 12.
[256] *See* Karen Steinhauser, *Everyone Is a Little Bit Biased*, AMER. BAR ASS'N BUS. L. TODAY (Mar. 16, 2020), https://www.americanbar.org/groups/business_law/publications/blt/2020/04/everyone-is-biased/; Jenée Desmond-Harris, *Implicit Bias Means We're All Probably at Least a Little Bit Racist*, VOX (last updated Aug. 15, 2016), https://www.vox.com/2014/12/26/7443979/racism-implicit-racial-bias.
[257] *Cf. supra* note 141 & accompanying text (discussing representativeness bias).
[258] *See* U.S. ENVIRONMENTAL PROTECTION AGENCY, RADIATION HEALTH EFFECTS (last updated Oct. 2, 2024), https://www.epa.gov/radiation/radiation-health-effects.

succinctly, "the higher the dose, the greater the risk."[259] So too do multiple exposures to these AI biases drive up the risks; likelihood of those risks being actualized; and the intensity and scope of the corresponding first, second, and third order impacts.

B. Bias Inheritance, and Inherited Bias

Bias inheritance is an AI bias mechanism by which, once some type of bias is injected into the lifecycle, the effects of that bias, absent detection and contravening measures, passes on through and affecting the remainder of that lifecycle. Thus, **inherited bias** is that downstream bias that is subsequently and passively injected into the AI enterprise lifecycle at one or more points after the initial bias arose or was injected and that is so as a consequence of that initial bias injection.[260] Much like poisoning a stream of water, the toxicity of that initial bias flows downstream, affecting the health of the stream and all that it touches or otherwise impacts.

The most obvious, incalcitrant, and unmitigated inherited AI biases are the systemic biases of inequality produced throughout society's long histories and still persisting in numerous and intersecting forms today. This type of societal and cohort bias contaminates much, if not virtually all, data about people used in AI systems if left unrecognized and unaddressed through corrective actions undertaken by AI creator-vendors and users. Thus, society gave and gives birth to biases that are, in turn, passed on, that is, inherited, throughout the AI lifecycle.

Thus, inherited bias is the progeny of its earlier-injected bias, or **parent bias**.[261] Consider the design and development of AI systems for facial recognition. Caucasian male faces predominate among the facial

---

[259] *Id.*

[260] This Article begins its conception of "inherited bias" as Hellström and his co-authors describe it from their machine learning bias literature survey. *See* Hellström et al., *supra* note 24, at 4. They categorize inherited bias as a data generation bias, but limit their discussion of this important phenomenon to bias having been "inherited" from the biased output of AI systems and "inherited" by AI systems that are subsequent in the computational chain of analysis. *See id.* As to the NIST report, note that it cites to Hellström and company and continues their too-narrowly-scoped conception of inherited bias. Without explanation, however, the NIST report categorizes inherited bias as a type of processing or validation bias. *See* NIST REP'T, *supra* note 30, at 52 (definition); *id.* at 8 (categorization).

As discussed in the text, this Article explains that the phenomenon and mechanism of bias inheritance and its progeny "inherited bias" are much more broadly operative and impactful in the AI lifecycle than envisioned in Hellström and his colleagues' article or the NIST report.

[261] Multiple AI biases may function as a parent bias. For ease of reading, this discussion refers to parent bias in the singular.

images datasets chosen to create the training dataset for such systems.[262] This parent bias is injected early in the data phase of the AI lifecycle. As cart follows horse, so too does progeny follow parent. Thus, the error caused by the parent bias in the data phase gives rise to inherited bias during the learning phase, the modeling phase, and so on throughout the lifecycle.[263]

C. Bias Amplification, and Amplification Bias

The mechanism of bias inheritance has its companion, a concept that is similar, but not identical to it: the mechanism of bias amplification.[264] To begin, the Article defines and distinguishes two terms with bias amplification being a mechanism by which AI biases work and amplification bias being a particular type of AI bias.[265] A specific form of error propagation,[266] **bias amplification** is the progressive exacerbation of existing biases as they persist within the human-machine enterprise. One circumstance in which bias amplification observably manifests, occurs when the distribution of an AI system's initially predicted results changes in comparison to the distribution of earlier such results.[267] Bias amplification may occur even

---

[262] *See generally* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, PROC. 1ST CONF. ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY, *in* PROC. OF MACHINE LEARNING RES. 77 (2018), https://proceedings.mlr.press/v81/buolamwini18a.html.
; *see also* P. Jonathon Phillips et al., *An Other-Race Effect for Face Recognition Algorithms*, 8 ACM TRANS. ON APPL. PERCEPT., Art. No. 14, 1-2, 7-10 (2011), https://dl.acm.org/doi/10.1145/1870076.1870082 (examining, comparing, and identifying that Eastern Asian- and Western-developed facial recognition algorithms perform less accurately, that is, in biased ways, in recognizing Caucasian- and Eastern Asian-appearing faces, respectively).
[263] *See* Loza de Siles, note 120, at 517; Michael L. Litman et al., G*athering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report* 54-55 (2021), https://ai100.stanford.edu/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence-ai100-2021-study.
[264] *See generally* Leino et al., *supra* note 187.
[265] Care is advised when studying these topics as some sources erroneously conflate or synonymize them.
[266] *See* NIST REP'T, *supra* note 30, at 51; *but see id*. at 49 (although without defining, identifying behavioral bias as one that, if not identified & addressed in data processes, may propagate to create learning bias or model bias or both). For more on behavioral bias discussed in an AI context, *see*, *e.g*., Lauren E. Willis, *Deception by Design*, 34 HARV. J. OF L. & TECH. 115 (2020).
[267] *See* Leino et al., *supra* note 187, at 1.

where there is no observable bias in those initial results, however.[268] In other instances, bias amplification may be unavoidable, meaning that it cannot be mitigated without negatively impacting the AI system's predictive accuracy.[269] Researchers, however, are positing ways in which certain instances of bias amplification may be identified and then mitigated.[270]

Corresponding to, but not synonymous with the AI bias amplification mechanism is **amplification bias**, a type of AI bias that is added to or compounded within the AI lifecycle as the result of bias(es) that earlier arose or were injected thereinto.[271] Amplification bias, therefore, represents the AI enterprise's additional or multiplicative degree of error.

To illustrate the concept of amplification bias, consider this hypothetical in which sampling bias[272] was earlier injected via the

---

[268] *See id.* Leino and colleagues go on to discuss instances in which bias amplification occurs, and that amplification may be identified, measured, and mitigated without impacting the accuracy of the predictor, that is, the ability of the AI system's classification model to produce the target. *See id.* at 5-8. In particular, they show that mitigable bias amplification occurs where: (1) the classifying predictor, *i.e.*, the AI system that predicts classifications, uses features that are weakly correlated with a desired predictive output; (2) the correlations of those features are "oriented" toward one particular class, *i.e.*, a so-called "majority" class, *id.* at 4-5; and (3) different classes, *e.g.*, Latinx, Caucasian, Black, Asian, Pacific Islander, and other racial, ethnic, or cultural groups, and, therefore, these features are represented in an imbalanced way, that is, exhibit non-parity or asymmetry, within the training dataset. *See id.*

[269] *See id.* at 1, 3-4.

[270] Leino and colleagues showed that such mitigable bias amplification occurs where: (1) the classifying predictor, *i.e.*, the AI system that predicts classifications, uses features that are weakly correlated with a desired predictive output; (2) the correlations of those features are "oriented" toward one particular class, *i.e.*, a so-called "majority" class, *id.* at 4-5; and (3) different classes, *e.g.*, Latinx, Caucasian, Black, Asian, Pacific Islander, and other racial, ethnic, or cultural groups, and, therefore, these features are represented in an imbalanced way, that is, exhibit non-parity or asymmetry, within the training dataset. *See id.*

These conditions result in such weakly correlated features being over-emphasized with the predictor. That over-emphasis, combined with the other circumstances described, *supra*, causes the classifier to poorly distinguish classes. *See id.* As the system iterates through the training dataset or in application to live data, the classification biases toward the majority class arise, if not already present, and subsequently increase in additive fashion. *See id.* Leino and co-authors christen this type of bias as "feature-wise bias" and identify it as a type of model bias. *Id.* at 1. They call the phenomenon by which such bias increases "feature-wise bias amplification" or, for short, "bias amplification." *Id.*

[271] As a point of care to be taken, note that the NIST Report cites to the work by Leino and colleagues discussed in this Section. It, however, transposes the terminology, calling the phenomenon "amplification bias," rather than bias amplification as discussed by Leino and colleagues, and categorizing it as a statistical or computational processing or validation bias without elaboration. *See* NIST REP'T, *supra* note 30, at 8, fig. 2 & 49.

[272] *See* Whittaker et al., *supra* note 128, at 8.

training dataset into the AI lifecycle. Take it that this sampling AI bias accounts for a bias score of 10 points, that is, a 10% error or deviation from the truth. The training dataset, along with its 10-point bias score, is ingested by the learning process, that is, within the learning span of the AI lifecycle. Imagine that the learning span of the lifecycle is entirely pristine. No new bias is injected into or originates in the learning process, the bias score of which, being discretely measured and with accuracy and precision, is zero (0-point). After the learning process is concluded and prior to any additional steps in the AI lifecycle, however, measurement reveals a 15% bias score. In this scenario, the bias score at the end of the data span of the lifecycle and prior to the start of the learning span was 10%, but at the end of the subsequent and entirely unbiased learning span, the bias score was 15%.

What has happened here? The sampling bias, which was injected by the training dataset, was amplified during the learning process. This hypothetical then reflects that, across the encompassed lifecycle spans, a 5-point amplification bias has arisen. This 5-point, or 5% delta, *i.e.*, difference, is between the dataset's bias score and the learning process' unparsed bias score. This delta of 5%, then, would constitute the amplification bias.[273]

To this point, the Article has discussed how AI biases are injected into the AI lifecycle at one or more phases and how those biases flow downstream, including with amplifying effect. The next subsection discusses ways in which AI biases interact causally with each other and others in which other AI bias mechanisms may additionally interoperate with these causal relationships.

D. Intercausality Between Biases and Interoperation of AI Bias Mechanisms

Bias begets bias. Causal relationships exist between biases. In her review of Virginia Eubanks' AUTOMATING INEQUALITY, Professor Dorothy Brown sketches a sequence of causal connections between biases.[274] Framed within the construct of this Article, that chain of causation runs as follows: There is an initiating social and cohort bias, specifically persistent systemic inequity bias in policing and over-policing poor, segregated, and black-concentrated neighborhoods. This initiating bias

[273] This analytical approach also could be useful in evaluating amplification bias where AI system outputs are subsequently used as inputs for other AI systems.
[274] Dorothy E. Roberts, *Digitizing the Carceral State,* 132 HARV. L. REV. 1695, 1719-21 (2019) (reviewing VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018)).

then produces arrest and other data in which poorer black people are over-selected as compared to people of other races and socioeconomic classes, resulting in data bias. This data bias produces observation bias,[275] which, in turn, produces confirmation bias.[276] This interlinked causal chain of biases is likened to a snake that eats its own tail in a toxic and self-perpetuating-fulfilling feedback loop.[277]

Although the causal chain may be neatly described in a linear fashion, intercausality between biases may be made more complex where causation interoperates with other bias mechanisms. In the following example, bias inheritance and intercausality may interoperate across a sequence of AI computations leading toward a target concept involving sentiment analysis.[278] To begin, the detection and analysis of smiles and other facial expressions and sentiment analysis may be used in depression and other mental health risk assessments and interventions. [279] Age biases may undermine the accuracy of facial expression analyses and, subsequently, sentiment analyses.[280]

Say that the first AI system in the sequence is a smile detection system the output of which exhibits an age bias.[281] These smile detection results, along with those as to other facial expressions, may feed into a

---

[275] As an illustration, *see* Jin et al., *supra* note 35, at § 1, para. 4. Jin and colleagues discuss the challenges and necessities in measuring dust quantities to correct for observation bias where storm images reflect dust and non-dust aerosolized particles. *See id.*

[276] *See* 2008 DICTIONARY, *supra* note 24, at 54 (defining confirmation bias as "[a] form of [bias] that may occur when evidence that supports one's preconceptions is evaluated differently from evidence that challenges these convictions").

[277] *See* EU AI Act, *supra* note 15, at Art. 15(4). Note that this Article, relying upon the technical literature, classifies feedback loop bias as a type of model bias. *See supra* note 200 & accompanying text. One of the challenges with the examination of AI biases is that multiple distinct biases may be known by the same name. *See*, *e.g.*, Hoffman & Podgurski, *supra* note 200, at 15-16 (discussing care algorithmically driven by demographic factors, rather than medical need).

[278] Bias amplification also may interoperate here.

[279] *See generally*, *e.g.*, Maya Mohan et al., *Depression Detection using Facial Expression and Sentiment Analysis*, IEEE PROC. 2021 ASIAN CONF. ON INNOVATION IN TECH. 1 (2021), doi: 10.1109/ASIANCON51346.2021.9544819. Physiologic parameters and other indicators also may be used. *See generally* Emile Loza de Siles, Military Application of Smart Garments for Stress Telemetry (July 8, 2018) (on file with author) (Harvard graduate data science course paper); Sergío Torres, Overview of a Revised Standard of Care Adaptable to the Advent of Emotion-Sensing Devices in Behavioral Health Pracitces (Aug. 13, 2022) (Artificial Intelligence & Social Justice course paper on file with author).

[280] *See generally* Hyungjoo Park et al., *Facial Emotion Recognition Analysis Based on Age-Biased Data*, 12 APPL. SCI. 7992 (Aug. 10, 2022), https://doi.org/10.3390/app12167992.

[281] *See* Hellström et al., *supra* note 24, at 4.

larger sentiment analysis[282] or other affective computing application.[283] Biases as to age, at once a societal and cohort bias and a data bias, in the smile detection system move through that AI lifecycle to produce biased results. The biases in the results from the first AI system, in turn, are inherited by the second system, here, as data bias(es) in the input for this second system and those, left undetected and unaddressed, similarly impact throughout that AI lifecycle to generate biased sentiment analysis results. In summary, the biases in the first system simultaneously constitute biases inherited by and caused by other AI biases in the second system.

The complexity with which AI biases interdigitate with one another across the AI human-machine life cycle or multiples thereof makes causal relationships between them exceedingly difficult to unravel.[284] As to proof of facts, novel explorations of attribution science, such as applied climate change causation,[285] should be undertaken in application to this complex world of AI biases and, ultimately, at which loci within AI lifecycles AI bias causation may be disrupted, governed, and human accountability identified and assigned. Until then, governance requirements should be laid out against those lifecycles and presumptions of causation adopted into the law.[286]

CONCLUSION

The Article aims to deliver its readers to a higher place in the ascent toward AI systems and uses that are free from AI biases or, at least, toward awareness of and work toward eliminating or mitigating AI biases and the harms that they may and do produce. With its conception of AI as a human-machine enterprise, its comprehensive lifecycle of processes in AI as that enterprise, its taxonomy and beginning compendium of fifty

---

[282] *See*, *e.g.*, Yuhao Kang et al., *Extracting Human Emotions at Different Places Based on Facial Expressions and Spatial Clustering Analysis*, ARXIV 1, 2 (May 7, 2019), https://arxiv.org/pdf/1905.01817.pdf.
[283] *See generally* ROSALIND W. PICKARD, AFFECTIVE COMPUTING (1997).
[284] *See*, *e.g.*, Tommasi, et al., *supra* note 138, at 3; *but compare* Patent '989, *supra* note 133, at 22, col. 20 (Claim 3) ("XAI models may also enable the creation of explanation path-traces for each of the input features in the underlying datasets, further assisting in the identification of potential bias being learnt in the minority and majority classes, in real-time.").
[285] *See generally* Michael Burger, Jessica Wentz & Radley Horton, *The Law and Science of Climate Change Attribution*, 45 COLUM. J. ENVTL. L. 56 (2020); Renée Cho, *Attribution Science: Linking Climate Change to Extreme Weather*, STATE OF THE PLANET (Columbia Univ., Oct. 4, 2021), https://news.climate.columbia.edu/2021/10/04/attribution-science-linking-climate-change-to-extreme-weather/.
[286] *Accord*, *e.g.*, Ajunwa, *supra* note 114, at 1726-34.

AI biases, the Article aims to have brought them a good way toward the knowing to which Deming exhorts those who want to control AI biases and toward establishing an informed governance to protect people and, consequently, communities, markets, civil society, and the rule of law from the impacts of those errors.

The Article with its novel systems and process control engineering approach aims to have set the stage for other interdisciplinary projects by which to critically examine and expand upon the ideas presented here and, together, to benefit AI law and policy debates and formulations. The operationalization of these ideas will be essential to their refinement and to evaluating and realizing their true import. Just as knowledge from engineering and many other disciplines has come into the law scholarship domain through this Article, so too, one hopes, knowledge from law and policy will come into those other disciplinary domains and into all organizations that create and use AI systems. The ultimate goal, and hope, for this work is that, by disaggregating AI biases and providing ordered actionable knowledge about them, the path toward the summit of informed and responsible AI governance and regulation is now clearer.